MBMG
Metabarcoding & Metagenomics

Metabarcoding and Metagenomics 10: 261–297 (2026)
DOI: 10.3897/mbmg.10.174450

Forum Paper

Guidance and best practices for species identification using eDNA metabarcoding – When do you call a cod a cod?

Alison W. Watts[1], Zachary Gold[2], Nastassia V. Patin[3], Nicolaus G. Adams[4], Jacoby D. Baker[5], Fouad El Baidouri[1], Erin Grey[6], Ann E. Holmes[7,8,9], Kym Jacobson[10], Sean Jungbluth[11], Tadashi Kajita[12], Anders Kiledal[13], Matthew A. Lemay[14], Enrique Montes[9,15], Frank E. Muller-Karger[16], Jeffrey Miller[1], Matthew B. Ogburn[17], Kathleen J. Pitz[5], Katherine Silliman[8,9], Andrew Thompson[18], Luke R. Thompson[8,9]

1 *University of New Hampshire, Durham, USA*
2 *Pacific Marine Environmental Laboratory, US National Oceanic and Atmospheric Administration, Seattle, USA*
3 *Scripps Institution of Oceanography, University of California, San Diego, USA*
4 *Northwest Fisheries Science Center, NOAA Fisheries, Seattle, USA*
5 *Monterey Bay Aquarium Research Institute, Moss Landing, USA*
6 *University of Maine, Orono, USA*
7 *University of California, Davis, USA*
8 *Northern Gulf Institute, Mississippi State University, Starkville, USA*
9 *Atlantic Oceanographic and Meteorological Laboratory, US National Oceanic and Atmospheric Administration, Miami, USA*
10 *Northwest Fisheries Science Center, US National Oceanic and Atmospheric Administration, Newport, USA*
11 *Estuary and Ocean Science Center, Tiburon, USA*
12 *University of the Ryukyus, Okinawa, Japan*
13 *University of Michigan, Ann Arbor, USA*
14 *Hakai Institute, Campbell River, Canada*
15 *Cooperative Institute for Marine and Atmospheric Studies, University of Miami, Miami, USA*
16 *College of Marine Science, University of South Florida, St Petersburg, USA*
17 *Smithsonian Environmental Research Center, Edgewater, USA*
18 *Southwest Fisheries Science Center, US National Oceanic and Atmospheric Administration, La Jolla, USA*

Corresponding author: Alison W. Watts (alison.watts@unh.edu)

Abstract

With the rapid uptake of environmental DNA (eDNA) metabarcoding for a wide range of conservation and management uses, there is a growing need for guidance and best practices for species identification. A range of methodological and interpretive decisions influences taxa assignment. Here we provide a review and perspective of pitfalls and issues that can impact accurate taxonomic assignment across the full eDNA metabarcoding workflow: primer selection, lab handling, bioinformatics, taxonomic assignment, and communication of results. This paper addresses the complexities and challenges of the metabarcoding-based species identification process, offering recommendations for robust workflows and effective communication of findings. Accurate interpretation and communication of eDNA metabarcoding results requires an acknowledgment of methodological limitations such as incomplete reference databases, contamination risks, ambiguous sequences, and detection biases. Thus, we argue that transparency of methods and limitations, alongside proactive alignment of decisions with project objectives are critical for the successful application of eDNA metabarcoding in conservation and management decisions. We provide guidance for developing protocols that support species identification from eDNA metabarcoding sequences, and give recommendations on communication strategies for stakeholders and end

OPEN ACCESS

Academic editor: Gert-Jan Jeunen
Received: 23 October 2025
Accepted: 1 March 2026
Published: 25 March 2026

Copyright: This is an open access article distributed under the terms of the CC0 Public Domain Dedication.

Citation: Watts AW, Gold Z, Patin NV, Adams NG, Baker JD, El Baidouri F, Grey E, Holmes AE, Jacobson K, Jungbluth S, Kajita T, Kiledal A, Lemay MA, Montes E, Muller-Karger FE, Miller J, Ogburn MB, Pitz KJ, Silliman K, Thompson A, Thompson LR (2026) Guidance and best practices for species identification using eDNA metabarcoding – When do you call a cod a cod? Metabarcoding and Metagenomics 10: e174450. https://doi.org/10.3897/mbmg.10.174450

users. To support these recommendations, we outline steps in the workflow that can impact species identification, with a discussion of the strengths and weaknesses at each stage. Ultimately, this guidance can improve the accuracy, reliability, and usability of eDNA and other metabarcoding and amplicon sequencing approaches to species identification while fostering trust and understanding among diverse end users.

Key words: Bioinformatics, ecosystem, environmental DNA, metabarcoding, monitoring, species identification

Introduction

The use of environmental DNA (eDNA) for species identification has increased rapidly in recent years. eDNA can provide a non-invasive and sensitive method for detecting species presence, even for rare or elusive organisms (Duarte et al. 2023; Takahashi et al. 2023; Bernatchez et al. 2024). Researchers and managers are increasingly adopting eDNA metabarcoding methods to survey biodiversity, monitor invasive species, assess ecosystem health, and inform conservation efforts. Advances in sequencing technology, bioinformatics tools, and reference databases have further facilitated the wider application of eDNA, making it a valuable tool across a range of ecological studies and management applications (Hakimzadeh et al. 2024; Çevik and Çevik 2025; Curto et al. 2025).

However, accurate sequence-based species identification is complicated by multiple factors, including incomplete reference databases, the inability of some primers to distinguish closely related species based on the target region, differing results based on quality control, dereplication, annotation workflows, and sensitivity to parameter choices such as confidence thresholds (Hleap et al. 2021; Keck et al. 2023). Decisions made at each step will change the level of specificity and sensitivity of the analysis. There is an increasing realization that standardization of steps in the eDNA workflows is key to transitioning eDNA to operational, mission-driven applications (Berry et al. 2021; Takahashi et al. 2025; Theroux et al. 2025). Careful documentation and reporting of all steps provides transparency and allows users to identify key decision points within the process. These decisions, and implications for use of the data, must be communicated carefully to stakeholders who may be less familiar with eDNA methods (Klymus et al. 2024; Stein et al. 2024).

We provide guidance for developing protocols that support species identification from eDNA metabarcoding sequences, and give recommendations on communication strategies for stakeholders and end users. To support these recommendations, we outline steps in the workflow that can impact species identification, with a discussion of the strengths and weaknesses at each stage. The overall goal of this paper is to synthesize the existing literature and highlight various pitfalls that may occur when interpreting and analyzing eDNA metabarcoding data.

The discussion and examples presented here are drawn from the experience of the authors, which is primarily in aquatic systems, but are broadly applicable to metabarcoding-based species identification in any taxonomic group. In more dense communities, such as microbial systems, the general guidelines

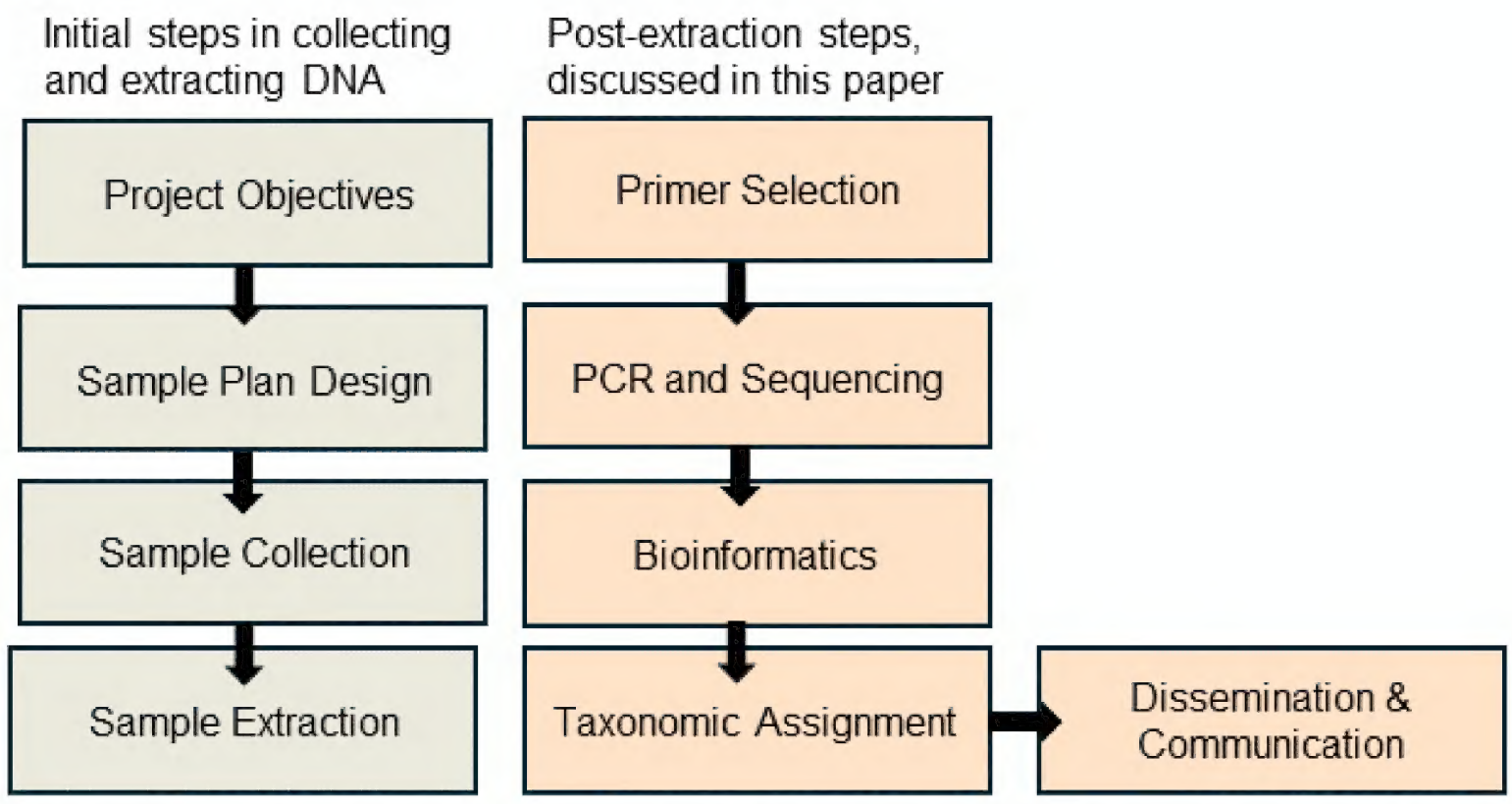

Figure 1. All steps in the metabarcoding workflow potentially affect taxonomic assignment and species identification. In this paper we focus on the post-extraction processes.

related to primer choice and lab protocols are still relevant, but species-specific identification is limited to a more targeted community.

An eDNA metabarcoding workflow begins with project objectives, sample plan design and sample collection, and then moves into extraction, polymerase chain reaction (PCR), sequencing, bioinformatics, and development of data products. This paper focuses on the effects of the post-extraction processes on species identification and interpretation, beginning with primer selection, PCR, sequencing, then bioinformatics and taxonomic assignment. Guidance presented here is applicable across metabarcoding workflows which could vary in taxonomic focus or sequencing strategy. In general, sample collection, preservation, and extraction choices govern the recovery of total and target DNA from the environment. For example, collecting more total DNA from the environment, increasing storage efficacy, and enhancing overall extraction efficiency improve eDNA metabarcoding broadly (Spens et al. 2017; Dickie et al. 2018; Sales et al. 2019; Peterson et al. 2022; Andruszkiewicz Allan et al. 2025). There are cases such as the inclusion of bead beating steps during extraction or targeting of different size fractions during sampling, that preferentially increase the recovery of some taxa as opposed to others, but generally the solution to this problem is to fundamentally better capture the full community (Anderson and Thompson 2022). Analytical choices made post-extraction are nuanced, and often have trade-offs that are difficult to navigate, especially for novice eDNA practitioners or users of the data. Our goal is to provide a primer on these often complex interrelated decisions to aid conservation and resource managers as well as early career scientists. Finally, we discuss dissemination of results and communication with stakeholders, including the difference between taxonomic assignment and species identification (Fig. 1).

Primer selection

Primer choice is the first and most fundamental factor influencing species identification in metabarcoding (e.g. Freeland 2017; Macher et al. 2023; Xu et al. 2024). Primer selection should be considered early in the project planning process, but at

any point, extracted samples may be re-analyzed using a different primer set(s) to meet new or revised project goals (Gold et al. 2022). Primers are synthetic oligonucleotides that bind to either end of a target DNA sequence to enable its PCR amplification. Metabarcoding primers are typically used to amplify sequences that are variable among species to maximize species discrimination (Fontes et al. 2024), but conserved at both ends to maximize primer binding and subsequent amplification across many species. The conflicting nature of these two aims, plus other considerations such as reference sequence availability (Ahmed et al. 2019), non-target sequence amplification (Collins et al. 2019), and amplicon length (Callahan et al. 2019; Srinivas et al. 2025), makes primer choice a complex decision that involves multiple trade-offs. In many cases, a literature search of similar studies in the study area targeting the same species can help identify an appropriate primer set(s).

When selecting primers, we recommend systematically assessing the strengths and weaknesses of candidate primer sets. First, the target species list should be co-developed with the end-user to ensure realistic and clear expectations up front. Candidate primers can then be found in the literature (Zhang et al. 2020) or through communication with other researchers. There are multiple primers developed for specific taxonomic groups that are widely used (e.g., Amaral-Zettler et al. 2009; Leray et al. 2013; Miya et al. 2015; Parada et al. 2016); in many cases, this review will identify appropriate primer set(s). However, small differences in target species may dictate specific primer choices; for example, *Gadus* (cod) species are difficult to distinguish using the common 12S MiFish (Miya et al. 2015) primer, but can be effectively resolved using the Gadid_ND1 primer (Ledger et al. 2024). Primer sets vary in specificity, ranging from broad categories such as fish or bacteria to species-specific targets. For example, *Sebastes* rockfish, *Microcystis* (Min et al. 2021; Kiledal et al. 2024; Ishii et al. 2025; Ledger et al. 2025; Marín et al. 2025). Primers with greater taxonomic breadth generally have less species-level specificity, highlighting a key trade-off in primer selection (So et al. 2020).

Additional validation can be done through a candidate primer assessment approach, of which there are two general options: *in vitro* and *in silico* testing (So et al. 2020; Espinosa Prieto et al. 2024). *In vitro* testing, which is metabarcoding of specimen tissue voucher-derived DNA extracts, is the most straightforward and accurate way to evaluate primers, but can be costly and challenging to implement if target species tissue or DNA is not readily available (Gold et al. 2021). *In silico* primer testing is more cost effective and less logistically challenging, making it the more popular choice (Boyer et al. 2016; So et al. 2020; Kalendar et al. 2024). *In silico* primer testing involves gathering reference sequences for target species [from a curated library like BOLD (Ratnasingham et al. 2024) or a public library like GenBank (Benson et al. 2015) and then evaluating primer chemistry and sequence alignments computationally. Published tools such as GAPeDNA (Marques et al. 2021), and MultiBarcodeTools (Zhu and Iwasaki 2023) can be used for development of *in silico* validation of metabarcoding primers. For example, GAPeDNA was used to assess the coverage and specificity of 21 published primer pairs, both alone and in combination with each other, for Great Lakes fish taxa (Mahon et al. 2022). Important metrics to consider during *in silico* evaluation include:

- Reference sequence availability: How many target species have an available reference sequence for this locus? This will set the limits of your

evaluation. For example, prior to a large sequencing effort only 53% of species actively monitored by long term research programs had reference sequences for the MiFish 12S locus compared to 100% for the Folmer COI locus (Hastings and Burton 2008; Gold et al. 2021).

- Amplification likelihood: Amplification likelihood of a target sequence by a given primer set can be estimated based on the number, location, and type of base-pair mismatches for given PCR conditions like primer concentration, annealing temperatures, and primer properties such as primer length or GC content. Exact matches between target and primer sequences indicate a high likelihood of amplification. Lower amplification likelihood will result in either reduced amplification rates (termed *primer bias*) or failure to amplify.
- Amplicon discriminatory power (e.g. Mahon et al. 2022; Fontes et al. 2024): To distinguish different species, their amplicon sequences must vary sufficiently. Creating a sequence dissimilarity matrix, multiple sequence alignment, or cluster-based sequence comparisons between all species' amplicon sequences can help assess this criterion. Primers can also capture intra- and inter-specific variation through haplotypic diversity. Metabarcoding primers can also be selected specifically for discriminating haplotypes between species (Adams et al. 2023; Serrana and Watanabe 2023).
- Amplicon length: While shorter amplicons are preferred for eDNA studies, especially if samples are expected to have degraded DNA, longer amplicons may improve amplicon discriminatory power (Buetas et al. 2024). The reliability and availability of sequencing technologies also influence this decision, with a need to align amplicon length with sequencing technology capabilities. Recent advances in long-read sequencing technologies have opened the door to longer targets that were previously inaccessible (Zhang et al. 2020), providing an avenue for enhanced taxonomic assignment from eDNA metabarcoding approaches.
- Non-target amplification: Some primers are more/less efficient at amplifying both target and non-target species, regardless of taxonomy (Collins et al. 2019). In some cases, amplification of non-target organisms may outcompete and overwhelm signals from the intended targets. Non-target amplification can be mitigated in the lab if target and non-target sequences are of significantly different lengths (Stoeckle et al. 2024a).
- Two or more primer sets: A multi-marker approach, including nested PCR, provides more robust data to support taxonomic identification, but the amount of lab, sequencing, and analysis effort required scales quickly (Lyet et al. 2021; Mahon et al. 2022; Ferreira et al. 2024).

Primer selection will be most effective when integrating information from literature reviews, regional expertise, and both *in vivo* and *in silico* validation (Beng and Corlett 2020). Once the primers have been selected and eDNA samples collected, samples are transported to, and processed in, a designated lab. In some cases, analysis will be conducted by the same team that conducted field work, while in others, samples may be sent to a commercial or collaborating lab (Vasselon et al. 2025). The recommendations below apply to either in-house or outsourced sample processing.

PCR and Sequencing

Samples must be handled with strict contamination control in the lab to ensure accurate species identification (Goldberg et al. 2016). This includes using sterile equipment, processing samples in dedicated clean areas, and including negative controls to detect potential contamination. Improper handling may result in cross-contamination between samples, contamination from external sources, or degraded DNA from poor storage. Failure to adhere to these best practices can lead to false positive detections due to cross-contamination or false negative results from degraded DNA. These errors can ultimately result in inaccurate species lists and misleading ecological interpretations (Darling et al. 2021).

Lab Contamination from Other Tissues/Controls

DNA contamination from non-target tissues or controls can result in false positives or skewed species abundance. Dedicated lab areas with unidirectional air flows, rigorous decontamination (UV, bleach, commercial solutions, etc.), the inclusion of negative controls (extraction and PCR blanks) in every batch, and utilization of closed versus open filtration (Patin and Goodwin 2023) reduces chances for contamination. Any high-template positive DNA control stored or present in the lab area should either be from taxa with no plausible ecological overlap (Stoeckle et al. 2022), or include a synthetic insert sequence (Sepulveda et al. 2020).

Degradation during handling

Prolonged storage, repeated freeze/thaw cycles, or improper handling can degrade DNA, particularly impacting the detection of rare species. Maintaining samples at –20 °C for short term, or –80 °C for long term storage, minimizing freeze/thaw cycles (create aliquots before first freeze - e.g., archive, stock, working stock), and regularly assessing DNA quality (have an *a priori* decision on when to throw out stocks, especially if stock is for quantification, such as a qPCR standard) can enhance DNA storage and quality assurance (Poulsen et al. 2021; Ogbanga et al. 2023).

PCR protocol variability

Even with a standardized primer set, differences in DNA extraction methods and PCR conditions can bias species amplification. For PCR, thermal cycling conditions, choice of polymerase, inclusion of albumin serum, and buffer conditions can increase/decrease primer specificity (e.g., high-fidelity polymerase, single annealing temperature vs. touchdown PCR) (Gohl et al. 2016; Clarke et al. 2017). Detection of large-scale trends and patterns may be less sensitive to specific PCR conditions, but identification of rare amplicon sequence variants (ASVs) or species is more sensitive to methodological differences (Shaffer et al. 2025). For example, El Baidouri et al. 2025, found that adding inhibition removal and an optimized PCR protocol increased both the number of fish species and ASV diversity. Protocols should be standardized and validated with replicates and internal standards, and detailed sufficiently to allow reproducibility (Shea et al. 2023; Klymus et al. 2024; Takahashi et al. 2025).

Sequencing

Sequencing protocols and methodological details will be driven by sequencing platforms. Several different sequencing platforms can be used for metabarcoding, each with trade-offs. Illumina instruments (like MiSeq, NextSeq, or NovaSeq) are used in many labs and support relatively low cost high-throughput short-read sequencing, though they are limited to shorter amplicons (≤500 bp). Other platforms, including Oxford Nanopore devices (e.g., MinION), PacBio, and Ion Torrent, can produce longer reads, although the sample throughput rate is lower, and the bioinformatics may be more challenging due to higher error rates and variable read lengths (Brandão-Dias et al. 2025). Sequencing depth guides the number of reads returned per sample; deeper sequencing (more reads) may be helpful in detecting rare taxa (Singer et al. 2019), where only a few copies of DNA are present in the sample, but the cost of processing generally increases as high read counts are requested. Aligning sequencing choices with primer selection and project objectives is critical to ensure that the generated data meet overall research goals.

Amplicon libraries are inherently low in sequence diversity, which can create challenges when sequencing on platforms such as Illumina MiSeq. High concentrations of the same base on the flow cell can mask less abundant bases during base calling and clustering, resulting in inaccurate base calls. This issue is compounded when using target regions that only amplify relatively low diversity groups (e.g., MiFish, Gadid_ND1 primer)(Kozich et al. 2013). The addition of PhiX (Illumina n.d.) during sequencing alleviates this issue by introducing highly diverse sequences into the run, increasing base calling accuracy and overall run performance. Genomic libraries can also be included in an amplicon sequencing run to contribute sequence diversity and reduce the amount of sequencing effort lost to PhiX control reads. With all 'low diversity' amplicon libraries, a spike of 10–20% PhiX is commonly recommended (Bourlat et al. 2016), although concentration differs across instruments, with NovaSeq and NextSeq instruments often requiring 30–40% (Illumina n.d.).

Bioinformatics

Bioinformatics provides the computational framework necessary to process, analyze, and interpret the large amounts of DNA sequence data generated by eDNA studies (Petit-Marty et al. 2023; Hakimzadeh et al. 2024; Dos Santos and Blabolil 2025). Through a combination of sequence quality control, error detection, taxonomic assignment, and statistical analysis, bioinformatic methods enable researchers to transform raw sequence reads into reliable information on species presence, community structure, and ecological patterns. Robust bioinformatic pipelines are essential to ensure accuracy and reproducibility of eDNA-based biodiversity assessments and facilitate comparisons across studies and ecosystems. Decisions are made throughout a pipeline that impact results in multiple ways. We do not provide recommendations for specific software or pipelines, but rather offer guidance on those aspects of the methods that are particularly important for species identification. We note that in some cases, a taxonomy-free approach may better serve project goals (e.g., if measuring overall biodiversity), and in these projects, taxonomic assignment guidelines are less relevant. But in cases where a primary goal requires identifying

a specific species or community, the bioinformatics pipeline is a crucial part of the species identification process. We generalize bioinformatic workflows into two main steps: 1) quality control and dereplication, and 2) taxonomic assignments and we discuss how decisions in each step can influence species identification from eDNA metabarcoding approaches.

Sequence quality control & dereplication

Quality control of sequence reads is a key step in any bioinformatic pipeline to remove errors that are generated by laboratory and sequencing processes (Zhu et al. 2024). Importantly, the inclusion and removal strategies of errors can lead to incorrect species identification and inaccurate assessments of biodiversity. Errors are first addressed by filtering and trimming low-quality or off-target reads from raw reads using sequence quality score thresholds (Fig. 2a). Quality filters apply models to base-call quality scores to estimate the likelihood of sequencing errors (Callahan et al. 2016; Singh and Wahengbam 2025). Threshold selection should be informed by previous experiments and sequencing run quality. Additional trimming of sequences is often conducted when removing primers, indices, and adapters, either using fixed-length or targeted sequence alignment-based approaches. Incomplete removal of indices, sequencing adapters, or primers can alter the resulting sequences, often due to the inclusion of degenerate bases. Thus, filtering and trimming choices can influence which sequences are retained for dereplication and taxonomic assignments, and which species are identified.

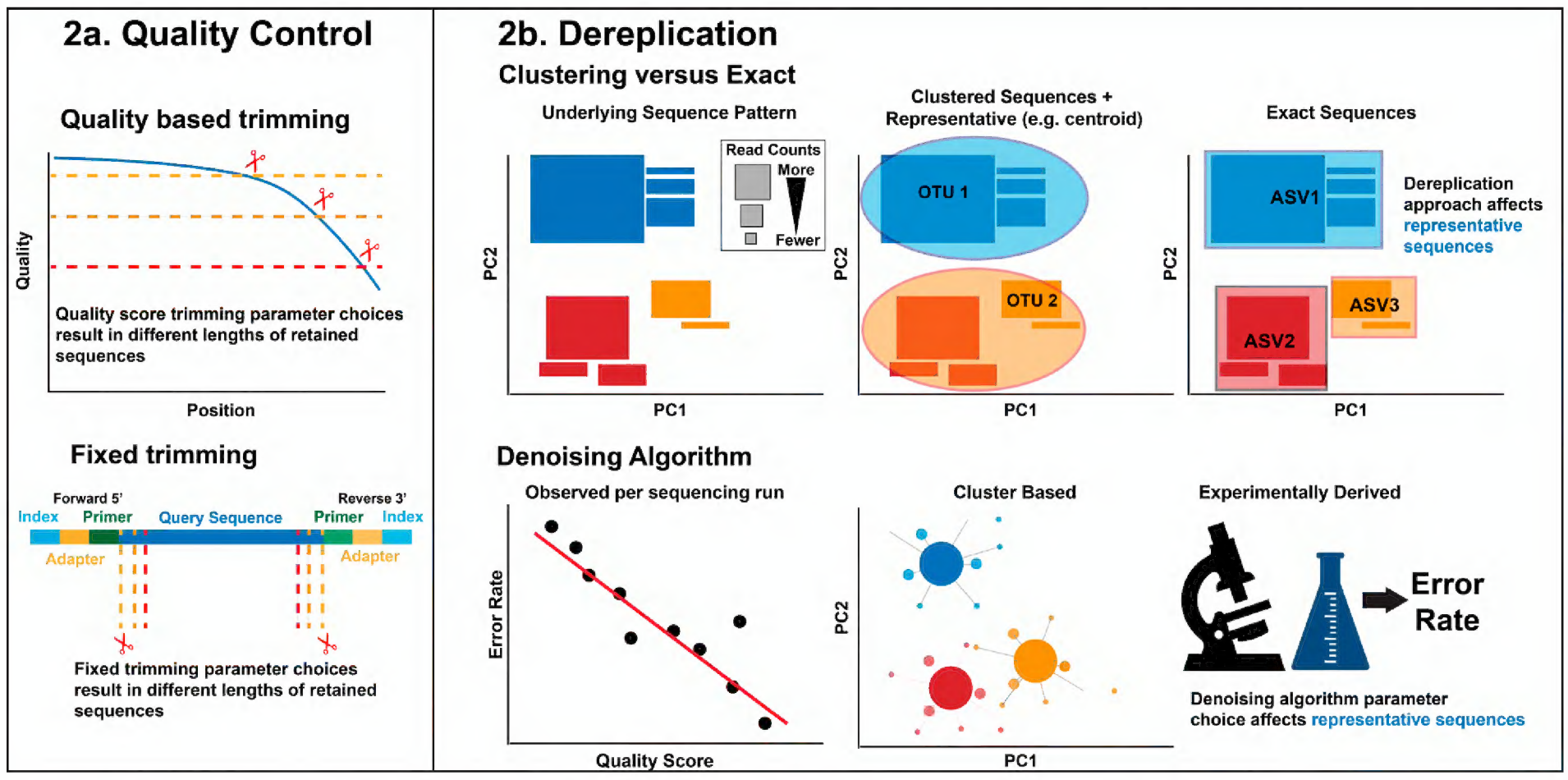

Figure 2. Quality control and dereplication influence underlying reference query sequences and thus have downstream implications on taxonomic assignments. Trimming approaches are either based on quality scores or fixed length, which can influence the length and accuracy of retained sequences (**2a**). Dereplication approaches also impact which sequences are retained from metabarcoding. There are two main dereplication strategies (**2b**), clustering by percent sequence identity (e.g., Operational Taxonomic Unit) or by exact or amplicon sequence variants (ESV/ASV). Denoising algorithms are used to indicate likely sequencing errors. The underlying algorithm and parameter selection can influence which sequences are retained with algorithms relying on either observed error parameters within each sequencing run, cluster, or similarity- based analyses of observed error parameters, or using fixed parameters derived experimentally.

Quality-controlled reads are then dereplicated to identify putative unique sequences or sequence clusters and reduce the computational load of redundant taxonomic comparisons. There are two main strategies for dereplicating amplicon sequence reads: cluster-based algorithms and denoising algorithms (Fig. 2b). Cluster-based algorithms rely on sequence similarity to generate Operational Taxonomic Units (OTUs), which can be subsequently curated with LULU (Frøslev et al. 2017), whereas denoising approaches use read frequency and sequence composition-based inference of likely sequencing errors to generate amplicon sequence variants (ASVs) (Quince et al. 2011). Choice of percent similarity thresholds in clustering algorithms greatly affects the resulting OTU composition and thus taxonomic assignments and estimates of biodiversity (Frøslev et al. 2017; Edgar 2018b). Likewise, algorithmic choices in denoising pipelines can influence the downstream ASVs retained. For example, DADA2 models the relationship between sequence error rates and quality score for each individual sequencing run, whereas Deblur uses a consistent set of error rates derived from experiments (Amir et al. 2017). Lastly, additional quality control steps in the dereplication process can be applied to remove sequencing errors (e.g. chimeric sequences). If sequencing errors pass the denoising step, they may be counted as a unique, low-frequency ASVs, and pervasive artifact ASVs can reduce the relative estimated abundance of taxa (Irber et al. 2024).

Together, algorithm and parameter choices in the quality control and dereplication steps affect which sequences are retained in the dataset, ultimately influencing which species are or are not detected from eDNA metabarcoding.

Dealing with ASVs: signal or noise?

The increasing use of ASVs instead of OTUs in eDNA metabarcoding has improved taxonomic resolution and reproducibility, but has also introduced new considerations for bioinformatic filtering and interpretation (Callahan et al. 2017; Chiarello et al. 2022). Unlike OTUs, ASVs retain exact nucleotide sequences, making it crucial to carefully separate true biological variants from sequencing artifacts. Rare ASVs, in particular, present challenges as they may represent genuine rare species or populations, but could also be PCR errors, sequencing errors, contamination, or technical artifacts like PCR chimeras or index-hopping (Huse et al. 2010; Edgar and Flyvbjerg 2015; Alberdi et al. 2018; Stoeckle et al. 2024b). Decisions regarding retention or exclusion of rare ASVs should thus balance sensitivity to real biological diversity against the specificity needed to prevent false-positive detections. Common strategies include applying abundance or prevalence thresholds (e.g., retaining ASVs above a relative abundance of 0.1%), requiring replicate confirmation, or employing occupancy modeling to evaluate detection consistency across technical replicates (Kelly et al. 2019; Gold et al. 2021). Incorporating post-denoising curation tools, such as the LULU algorithm, can further improve data quality by merging spurious sequences based on similarity and co-occurrence patterns without excessively removing rare, potentially informative taxa (Frøslev et al. 2017). Transparent documentation of these filtering criteria and rationales is essential to ensure reproducibility, facilitate cross-study comparisons, and clearly communicate the reliability of rare ASV detections.

One example from Gold et al. 2022 detected two species of cetaceans with a single detection out of four eDNA sample replicates and less than 100 reads

at the mouth of the Ports of Los Angeles and Long Beach, California. Here the authors suggest that these rare detections were likely the result of transport of eDNA from organisms outside the Port complex and demonstrate how utilization of site occupancy models can provide reliable and repeatable decision-making frameworks to disentangle true signal from noise in eDNA metabarcoding data. Given that both observed cetacean species are known to occur in the area, it is likely that this was a real signature, but given the low replication of the ASVs across the whole study (n = 1) and the management consequences of such species being present within an active Port complex, the authors excluded the detection from their statistical analyses and biological interpretations. Other researchers may have interpreted these observations differently, especially in the context of additional data (e.g. repeated observations over time). This highlights the need for transparent documentation of such decision making.

Taxonomic assignment

Accurate metabarcoding results rely on the detection of biologically valid sequences and the correct taxonomic assignment of those sequences. There are several methods of assigning taxonomy to amplicon sequences, with the most common being alignment-based methods (Hleap et al. 2021). Alignment-based methods, such as BLAST (Boratyn et al. 2013) and the "allpairs_global" function of vsearch (Rognes et al. 2016), compare sequences base-pair-by-base-pair to determine observed sequence similarity to reference sequences. Alignment 'quality' or score is based on both coverage of the query sequence against the reference sequence as well as the number of matching base pairs within the covered regions (Fig. 3A). Thresholds can then be applied to both of those parameters to establish cutoffs for accepted alignments. Such thresholds can be adjusted for different taxonomic ranks (e.g. species, genus, family) and should be based on the marker gene composition, marker gene inter- and intra-specific variability, marker gene length, and target taxonomic group. For example, a common threshold for species-level assignment of the 12S rRNA gene (MiFish region) to species level is 97–98% (Polanco et al. 2021; Sales et al. 2021). However, even within one marker gene and one target organism group, the identity threshold can vary widely among clades (Bokulich et al. 2018; Edgar 2018a; Curd et al. 2019; Gold et al. 2021; Polanco et al. 2021). If a sequence does not align to any reference sequence at or above the specified threshold level, indicating the observed query sequence is not represented in the reference database, a lower threshold (e.g., 95%) for generic level assignment can be used for assignment to a higher taxonomic level, such as genus or family (Galaska et al. 2023), although the appropriate threshold varies by marker, taxon, and project goal.

A second or complementary strategy to alignment-based taxonomic assignment is phylogenetic-based assignment. Here reference sequences are first aligned in a multiple sequence alignment and then used to construct large marker gene-specific phylogenetic trees (Fig. 3B). ASVs are then assigned based on their closest tree-based reference. Examples of phylogenetic-based classifiers include *tronko* (Pipes and Nielsen 2024), *EPA-NG* (Barbera et al. 2019), and *pplacer* (Matsen et al. 2010). Phylogenetic approaches have been shown to be particularly helpful for placing sequences that lack exact matches in reference sequence databases, as they provide confidence scores for every node of

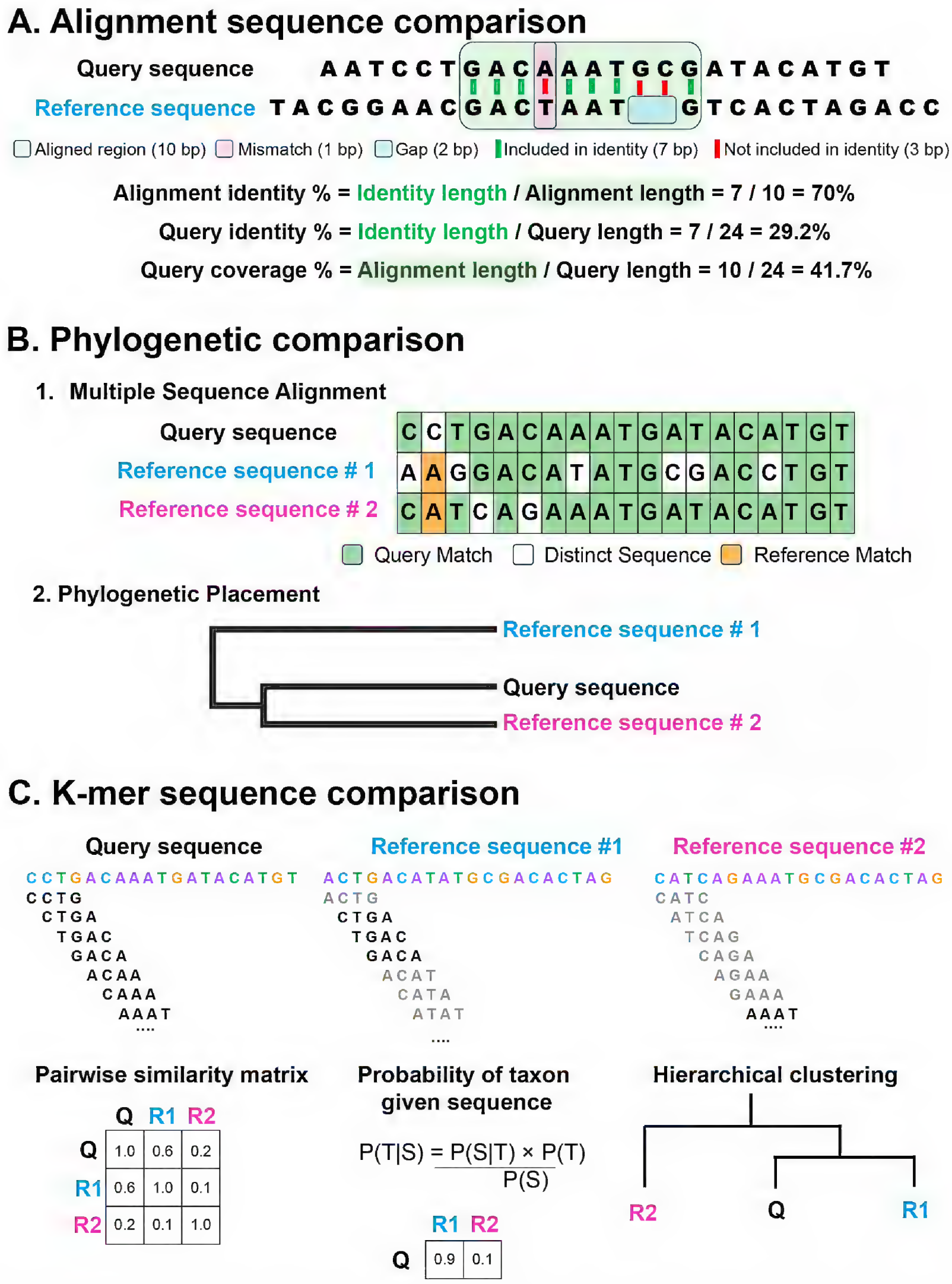
A. Alignment sequence comparison
Query sequence
A A T C C T G A C A A A T G C G A T A C A T G T
Reference sequence
T A C G G A A C G A C T A A T G T C A C T A G A C C
Aligned region (10 bp)
Mismatch (1 bp)
Gap (2 bp)
Included in identity (7 bp)
Not included in identity (3 bp)
Alignment identity % = Identity length / Alignment length = 7 / 10 = 70%
Query identity % = Identity length / Query length = 7 / 24 = 29.2%
Query coverage % = Alignment length / Query length = 10 / 24 = 41.7%
B. Phylogenetic comparison
1. Multiple Sequence Alignment
Query sequence
Reference sequence # 1
Reference sequence # 2
C C T G A C A A A T G A T A C A T G T
A A G G A C A T A T G C G A C C T G T
C A T C A G A A A T G A T A C A T G T
Query Match
Distinct Sequence
Reference Match
2. Phylogenetic Placement
Reference sequence # 1
Query sequence
Reference sequence # 2
C. K-mer sequence comparison
Query sequence
Reference sequence #1
Reference sequence #2
C C T G A C A A A T G A T A C A T G T
C C T G
C T G A
T G A C
G A C A
A C A A
C A A A
A A A T
A C T G A C A T A T G C G A C A C T A G
A C T G
C T G A
T G A C
G A C A
A C A T
C A T A
A T A T
C A T C A G A A A T G C G A C A C T A G
C A T C
A T C A
T C A G
C A G A
A G A A
G A A A
A A A T
Pairwise similarity matrix
Q R1 R2
Q 1.0 0.6 0.2
R1 0.6 1.0 0.1
R2 0.2 0.1 1.0
Probability of taxon given sequence
P(T|S) = P(S|T) × P(T) / P(S)
R1 R2
Q 0.9 0.1
Hierarchical clustering
R2
Q
R1

Figure 3. Three approaches to taxonomic assignment using reference sequences of known taxa. **A.** Alignment-based methods, which perform pairwise alignments and score the similarities based on nucleotide matches/mismatches and insertions/deletions. The highest-scoring pairwise alignment provides the assigned taxonomy; **B.** Phylogenetic based methods, which use multiple sequence alignments to generate phylogenetic trees. Taxonomy is assigned based on the closest evolutionary relationship to reference sequences. Final taxonomic assignments from these methods can be made through consensus or lowest common ancestor approaches whose choices affect taxonomic assignments; **C.** *k*-mer based methods compare sequence composition using short strings of nucleotides of length *k*. In this example, all black *k*-mers are present in the reference sequence, and query sequences with more black *k*-mers in a query are therefore more similar to the reference. *k*-mer frequency profiles can be linked to specific taxonomic groups by generating similarity matrices or building machine learning classifiers from reference sequence databases, which are then used to assign taxonomy to the highest scoring reference sequence.

the tree. In general, more conserved genes with reliable rates of mutation over evolutionary time will produce more accurate phylogenetic placements, as will markers with lower rates of divergence and longer sequences (Dunthorn et al. 2015). In addition, the methods used to build multiple sequence alignments and conduct phylogenetic placement will influence taxonomic assignments.

A major challenge of implementing phylogenetic-based approaches is the computational effort required to develop robust alignments of reference databases, particularly for marker genes like COI and 16S v4, which contain hundreds of millions of sequences in publicly available databases (Curd et al. 2024). However, recent advancements have leveraged advances in fast approximate likelihood approaches to assemble such trees (Pipes and Nielsen 2024), opening the door to more widespread adoption.

Lastly, *k*-mer based methods compare the similarity of small strings of nucleotides with length *k* ("*k*-mers") to determine sequence identity. They include naïve Bayes algorithms like the Ribosomal Database Project (RDP) classifier (Wang et al. 2007) and the scikit-learn naïve Bayes machine-learning classifier used in QIIME2 (Bokulich et al. 2018), as well as other programs like Kraken2 (Wood and Salzberg 2014; Lu and Salzberg 2020) and sourmash (Irber et al. 2024). Initially, alignment-free approaches to taxonomy assignment are likely favored due to their far greater computational efficiency. Often, but not always, k-mer based classification appears to be more accurate than alignment based methods, perhaps due to sequence misalignments and algorithmic intolerances (Bokulich et al. 2018; Gamage et al. 2019; Hleap et al. 2021; Mathon et al. 2021). However, similar to alignment and phylogenetic methods, classifications errors are influenced by the algorithm parameter choices and the composition/completeness of the reference training database (Vinje et al. 2015; Robeson et al. 2021). Parameterization of either algorithm will vary depending on the acceptable rate of false-positive and false-negative classifications of the application, and classifiers perform best when the sequences discovered in the samples are represented in the reference training dataset (Bokulich et al. 2018). For example, the accuracy of the k-mer based naïve Bayes classifier can be improved further by including prior knowledge of species distributions and/or abundances (Bokulich et al. 2018). As such, if there are multiple hits to known taxa, the classification will be weighted toward the presumed more ecologically valid species (Kaehler et al. 2019; Gold et al. 2021; Mugnai et al. 2023).

The choice of parameter values for each algorithm can heavily influence the final taxonomic assignments of each sequence (Bokulich et al. 2018). The choices for these parameters are particularly critical for the sequences that do not uniquely align to the reference database entries. For alignment and phylogenetic taxonomic assignment methods, parameters such as marker gene length, query coverage, and percent identity are critical (Fig. 4), while for *k*-mer based classifiers *k*-mer size and confidence threshold should be critically evaluated (Curd et al. 2019). The "consensus top hit method" stage of alignment-based methods selects a user-designated number of reference sequence hits over a user-designated percent identity. In instances where the taxonomy conflicts among the top hits, sequences are classified to the lowest taxonomic rank (a Lowest Common Ancestor or LCA method) that is shared among the top hits. Setting thresholds too high or too low can result in a high proportion of unclassified and incorrectly classified sequences.

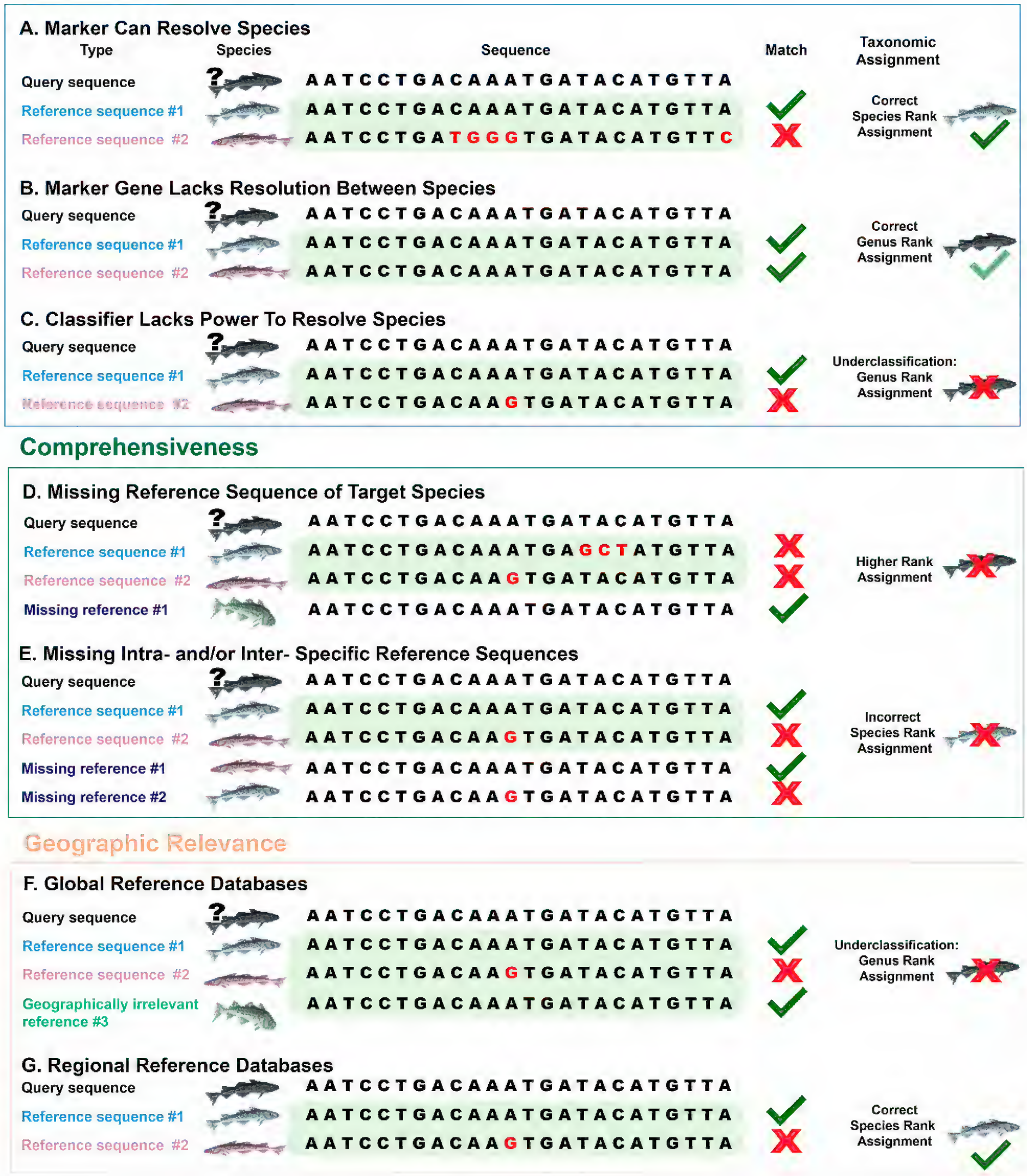
Resolution
A. Marker Can Resolve Species
Type
Species
Sequence
Match
Taxonomic Assignment
Query sequence
Reference sequence #1
Reference sequence #2
A A T C C T G A C A A A T G A T A C A T G T T A
A A T C C T G A C A A A T G A T A C A T G T T A
A A T C C T G A T G G G T G A T A C A T G T T C
Correct Species Rank Assignment
B. Marker Gene Lacks Resolution Between Species
Query sequence
Reference sequence #1
Reference sequence #2
A A T C C T G A C A A A T G A T A C A T G T T A
A A T C C T G A C A A A T G A T A C A T G T T A
A A T C C T G A C A A A T G A T A C A T G T T A
Correct Genus Rank Assignment
C. Classifier Lacks Power To Resolve Species
Query sequence
Reference sequence #1
Reference sequence #2
A A T C C T G A C A A A T G A T A C A T G T T A
A A T C C T G A C A A A T G A T A C A T G T T A
A A T C C T G A C A A G T G A T A C A T G T T A
Underclassification: Genus Rank Assignment
Comprehensiveness
D. Missing Reference Sequence of Target Species
Query sequence
Reference sequence #1
Reference sequence #2
Missing reference #1
A A T C C T G A C A A A T G A T A C A T G T T A
A A T C C T G A C A A A T G A G C T A T G T T A
A A T C C T G A C A A G T G A T A C A T G T T A
A A T C C T G A C A A A T G A T A C A T G T T A
Higher Rank Assignment
E. Missing Intra- and/or Inter- Specific Reference Sequences
Query sequence
Reference sequence #1
Reference sequence #2
Missing reference #1
Missing reference #2
A A T C C T G A C A A A T G A T A C A T G T T A
A A T C C T G A C A A A T G A T A C A T G T T A
A A T C C T G A C A A G T G A T A C A T G T T A
A A T C C T G A C A A A T G A T A C A T G T T A
A A T C C T G A C A A G T G A T A C A T G T T A
Incorrect Species Rank Assignment
Geographic Relevance
F. Global Reference Databases
Query sequence
Reference sequence #1
Reference sequence #2
Geographically irrelevant reference #3
A A T C C T G A C A A A T G A T A C A T G T T A
A A T C C T G A C A A A T G A T A C A T G T T A
A A T C C T G A C A A G T G A T A C A T G T T A
A A T C C T G A C A A A T G A T A C A T G T T A
Underclassification: Genus Rank Assignment
G. Regional Reference Databases
Query sequence
Reference sequence #1
Reference sequence #2
A A T C C T G A C A A A T G A T A C A T G T T A
A A T C C T G A C A A A T G A T A C A T G T T A
A A T C C T G A C A A G T G A T A C A T G T T A
Correct Species Rank Assignment

Figure 4. Key taxonomic classification issues for reference databases. Issues include **A–C.** resolution, the ability of known markers that can distinguish among target and non-target taxa; **D, E.** comprehensiveness, the presence of diverse potential species and congeners and all possible ASVs (amplicon sequence variants) within each taxon, accounting for population genetic variation at markers; **F, G.** geographic relevance, whether the database includes species known to be found in your region or habitat, and **H, I.** taxonomic curation whether the database includes accurate species labels including up to date species nomenclature and trusted sequence annotations.

Phylogenetic-based assignment with the LCA approach may have an advantage in this regard, as it is not limited by the vagaries of taxonomic rank delineations (e.g., lumping versus splitting rates across genera and families),

enabling the placement of query sequences among a set of references, often facilitating higher resolution at sub-rank levels.

Previous work has found that both sequence alignment and *k*-mer based approaches work particularly well with complete, comprehensive reference databases with parameter decisions having just as strong of an influence on taxonomic assignment accuracy as algorithm choice (Bokulich et al. 2018; Edgar 2018a; Curd et al. 2019; Gold et al. 2021; Hleap et al. 2021; Mathon et al. 2021). In contrast, phylogenetic placement strategies work well when conducting taxonomic assignments with sparse reference databases, although they are more computationally intensive (e.g. marine invertebrates) (Pipes and Nielsen 2024).

Together, classifier and parameter choices can have significant impact on which species are identified from eDNA metabarcoding approaches. Researchers should take care to conduct validation and benchmarking studies to confirm the efficacy of taxonomic assignment approaches.

Reference databases

Reference databases contain sequences used for the identification of species. They range from large but poorly annotated public repositories (e.g., GenBank), to highly specialized 'bespoke' databases valid only for particular taxa at particular locations (e.g., FishCARD, (Gold et al. 2021). Examples of current databases are included in supplementary materials (Suppl. material 1). Here we define four general types of reference databases: 1) comprehensive, uncurated public sequence repositories, 2) expert-curated reference databases, 3) uncurated, marker-specific reference databases, and 4) custom, curated reference databases. Comprehensive, uncurated public sequence repositories such as NCBI GenBank, ENA, and DDBJ serve as "parking lots" for user-submitted sequences, with the trustworthiness of sequences dependent on the accuracy of the user submissions. Although these public sequence repositories may contain errors (Leray et al. 2013; Schoch et al. 2020; Keck et al. 2023; Chorlton 2024; Renner et al. 2024), they remain highly used and comprehensive due to widespread international use by the research community.

To address quality concerns, researchers, institutions, and consortia have developed expertly curated reference databases with staff and resources dedicated to regular maintenance. These databases can have different communities of practice and levels of curation, and frequently include taxonomic curation, restriction to voucher specimens, and *in silico* sequence validation and removal of putative outliers. Many curated databases focus on either full mitogenomes or locus-specific nuclear genomes (e.g., BOLD, SILVA, GreenGenes), or target specific taxonomic groups, such as diatoms (Diat.barcode, rbcl), protists (PR2/18s), fish (MitoFish/12s), fungi (UNITE), and prokaryotes (SILVA) (DeSantis et al. 2006; Guillou et al. 2013; Quast et al. 2013; Miya et al. 2015; Rimet et al. 2019; Abarenkov et al. 2024; Ratnasingham et al. 2024). While typically derived from public repositories prior to curation, some accept direct submissions, potentially creating sequences exclusive to that curated database (e.g., BOLD).

Increasingly, software tools are enabling researchers to develop *marker specific databases* for diverse loci and target taxa in eDNA metabarcoding. These databases are typically marker-specific and range from uncurated (CRABS,

Metataxa2) to partially curated (RESCRIPt, rCRUX) (Bengtsson-Palme et al. 2018; Robeson et al. 2021; Jeunen et al. 2023b; Curd et al. 2024).

Each of the previously described types of databases can be further refined into custom, curated reference databases that may be specific to a single gene and geographic region of interest. For example, the FishCARD database includes only fish references from MiFish 12S sequences from the California Current Large Marine Ecosystem, based on a comprehensive list of regional fish species developed through an extensive literature review. Similarly, the Chesapeake Bay Barcode Initiative leverages high-quality voucher specimens of regional invertebrate and fish species to add new marker sequences as new primer sets are developed. As bespoke reference databases are often generated for a specific project, they can sometimes lack dedicated support for long-term maintenance and can rapidly become outdated due to taxonomy updates and newly generated sequences.

We define taxonomic curation as the curation of the taxa and taxonomic hierarchy being used by a particular reference database. Different reference databases utilize distinct taxonomic backbones. For example, GenBank uses the NCBI taxonomic backbone, PR2 employs its own 9-level hierarchy, SILVA utilizes the Genome Taxonomy Database (GTDB), and UNITE relies on a sequence similarity-based taxonomy (Parks et al. 2018, 2025). The taxonomic backbones of expert-curated databases such as UNITE, PR2, and SILVA are informed by the expertise of their communities of practice, whereas NCBI Taxonomy is an amalgamation of user input that often includes outdated and unaccepted taxonomic names. We highlight that these taxonomic hierarchies are distinct, though often overlapping, with taxonomic backbones used by global biodiversity platforms that leverage the Aphia platform to integrate taxonomic information across Global, Regional, and Thematic Species databases (Vandepitte et al. 2015). Efforts are underway to better unify taxa across the various taxonomic hierarchies (e.g., Biodiversity Information Standards TDWG). However, researchers today are challenged with navigating multiple existing taxonomic hierarchies. This results in new challenges as eDNA metabarcoding efforts increasingly use multiple marker loci simultaneously and thus may need to use multiple reference databases with different taxonomic hierarchies on the same sample. The existence of multiple taxonomic hierarchies across multiple loci makes synthesis of such datasets challenging.

Here, we highlight the need to acknowledge the advantages and disadvantages that a specific reference database or taxonomic hierarchy being used may have. We recognize that changing taxonomic backbones, along with increasingly expanded reference databases, present a monumental challenge for eDNA metabarcoding analysis and providing taxonomic assignments that follow Findable, Accessible, Interoperable, and Reproducible (FAIR) principles, as such classifications are not static and subject to change over time (Takahashi et al. 2025). Retaining and depositing representative OTU or ASV sequences is critical for updating taxonomic classifications as databases are updated.

Evaluating reference databases for trustworthiness is the key to enhancing the success and confidence of eDNA-based taxonomic assignments. Existing tools like GapDNA can be used to conduct gap analyses to evaluate the comprehensiveness of a given reference database (Marques et al. 2021). Likewise, tools like TAXXI and tax-credit can be used to evaluate resolution for a given

marker set (Bokulich et al. 2017; Edgar 2018a). Previous work has found that optimal parameters differ across marker loci (Edgar 2018b; Curd et al. 2019), highlighting the need for optimization for a given marker of interest to maximize taxonomic classification accuracy. Similarly, such optimization efforts can be used to understand resolution and confidence tradeoffs across types of classification errors (e.g. misclassification, overclassification, underclassification) (Edgar 2018a; Gold et al. 2021), which is important for informing value judgement conversations on which error types are more or less acceptable for a given eDNA metabarcoding project. To date, there are few tools to conduct systematic taxonomic curation of reference databases, with much of this work led by taxonomic experts that produce professionally curated reference databases like PR2 for protists, SILVA for microbes, FITS for fungi, and MitoFish for fishes. However, not all markers have dedicated research teams conducting taxonomic curation efforts and thus, researchers may need to curate their own reference databases appropriately (Nilsson et al. 2009; Guillou et al. 2013; Iwasaki et al. 2013; Quast et al. 2013; Bucklin et al. 2016; Jeunen et al. 2023b; Curd et al. 2024; O'Brien et al. 2024).

In addition, the geographic relevance of reference sequences is often completed manually and requires extensive knowledge of the existing taxa within a biogeographic region. Determining geographic relevance of observed sequences and species is better suited for well-studied areas with comprehensive species inventories (e.g., fishes of the California Current Large Marine Ecosystem, (Gold et al. 2021); fish and invertebrates of Chesapeake Bay; freshwater fishes in Switzerland, (Blackman et al. 2023)). Public global biodiversity observation databases, such as the Ocean Biodiversity Information System (OBIS) and the Global Biodiversity Information Facility (GBIF) can be used to build locally comprehensive, bespoke reference databases (Anon 2020; O'Brien et al. 2024; Anon 2026). However, similar to international sequencing repositories, these biodiversity databases have their own issues that may affect the trustworthiness of such biodiversity observations (Zizka et al. 2020). GBIF and OBIS also rely on different taxonomic hierarchies than GenBank or PR2 and thus require cross-walking of taxonomic references to accurately facilitate geographic curation. To date, no systematic tools exist to facilitate rapid geographic curation of species identifications from eDNA metabarcoding, highlighting the value of incorporating ecological and biogeographic expertise from regional and taxon-specific professionals in accurate interpretation of eDNA data.

It is also important to consider tradeoffs associated with using custom reference databases and post-classification processing versus expertly curated reference databases. Using the latter comes with a suite of advantages, as such databases often have been taxonomically curated, evaluated for their resolution, and systematically benchmarked. However, previous work has demonstrated the advantages of using custom reference databases that have been curated for geographic relevance (Gold et al. 2021; Mugnai et al. 2023; Pagenkopp Lohan et al. 2023) and have enhanced comprehensiveness (Arranz et al. 2020; Jeunen et al. 2023a; Curd et al. 2024).

Reference databases are often subsetted for various reasons like speeding up ASV classification procedures or reducing the appearance of ASV misclassification. To reduce the reference databases' size, full-length sequences and online repositories can be scanned with a particular primer pair to return

only those taxa (and their amplicon sequences) that could (theoretically) be detected (Bolyen et al. 2019; 2, Curd et al. 2024). In addition, it is important to dereplicate reference databases, especially for cluster-based classifiers, as algorithms can be biased toward overrepresented taxa. Dereplication efforts can also help identify marker resolution (e.g., which taxa have unique sequences). Example reference databases and their appropriate backups are included in Supplemental Information (Suppl. material 1).

Collectively, different reference databases can impact the final output of species generated from sequence data (Richardson et al. 2017; Hleap et al. 2021; Van Etten et al. 2023), but the exact outcomes vary based on target organisms, marker gene, project goals, and software parameters. For example, database coverage for fishes may be very high for a specific study region, whereas for others (and many other taxa; e.g., microbes, plankton) there will be an unknown large number of taxa missing from the database and unresolved issues in taxonomic classification. We therefore hesitate to recommend a single "best practice" or database for taxonomic assignment. Instead, researchers should approach taxonomic assignments with caution and conduct systematic evaluations and rigorous benchmarking of the performance of taxonomic assignments. Species-level assignments should be critically evaluated, especially when high confidence is required for a detection, such as regulatory and management implications (Kelly et al. 2024).

We encourage more widespread *in vitro* and *in silico* benchmarking of marker gene sequence taxonomic resolution, when project resources allow. Comparisons of assignment methods and optimization of parameters can be performed on a chosen reference database using tools such as TAXXI (Edgar 2018a), tax-credit (Bokulich et al. 2018), or the QIIME2 plugin RESCRIPt (Robeson et al. 2021). Previous work has found that optimal parameters are different across metabarcoding loci (Edgar 2018a). Mock communities relevant to target organisms can also be used as benchmarks for determining an optimal approach. In most cases, results can be validated to some degree using local ecological knowledge, different data streams (e.g., visual observations), or manual checks (e.g., manual sequence alignment with reference sequences of the annotated species). We suggest performing some or all of these validations when reasonable, but especially when results may be used for management or policy decisions that require high confidence and transparency. In all cases, each step of the analysis and all associated code should be reported for transparency and reproducibility.

Validating results

Identified species should be reviewed against expected and likely organisms in the region. This can be done during the sequence data processing as described above, using either a custom reference database curated for local species or a taxonomic assignment method that preferentially weights assignment of local species (e.g. Kaehler et al. 2019). However, it is important to consider the presence of unexpected taxa. If a detected species is not expected to be present in the system, several possibilities should be considered: 1) new detection of an organism not previously known to be living in the region, indicating previously unrecognized biodiversity or newly introduced species (e.g., invasive species, range shifts ((Pagenkopp Lohan et al. 2023); 2) detection of exogenous eDNA

sources (e.g., wastewater, food processing facilities, ballast water, (Darling et al. 2021)); or 3) process-based erroneous detections arising from either low level contamination, sequencing artifacts, ambiguous sequences, or incomplete reference databases. Each of these is discussed in more detail below.

1) New detection of an organism, invasive species, or rare species

In order to determine whether a detection of an organism is new to a region, a comprehensive assessment of what is known to live in the region is needed. Long-term monitoring programs, peer-reviewed literature, agency reports, and local experts can be used to generate lists of expected species. Local and regional resources are often more reliable than national databases. Resources include OBIS, GBIF, Fishbase, USGS, and local agencies (e.g., state website). The expected species should include potential invasives and range expansions. A list of expected species (both native and non-native) can be generated using the local and regional resources above. Regional watchlists (e.g., USGS Nonindigenous Aquatic Species information resource, USDA Invasive Species Profiles List) can be used to screen for unexpected invasive species.

Invasive aquatic species are inherently a regional issue because non-native species always have a native range. This is especially true for the freshwater realm, which is more clearly partitioned. Invasive marine species, especially invertebrates, can be widely distributed, and some have uncertain origins (cryptogenic). Researchers and managers should be aware of the potential for erroneous detections of AIS and native species when reference sequences are not available or closely related species share identical barcode sequences. Taxonomic assignments of invasive, endangered, and otherwise protected or listed species can have severe management implications and thus should be associated with greater confidence requirements. As Carl Sagan noted, “Extraordinary claims require extraordinary evidence” (Tressoldi 2011). Thus, researchers should conduct due diligence in the taxonomic assignments of invasive and listed species, including conducting multiple different assignment approaches and evaluating the completeness and resolution of reference databases for such taxa and their sister species.

2) Exogenous eDNA Sources

Exogenous eDNA detection herein refers to DNA that is present in a sample, but likely does not represent a species living in the immediate sampled environment. These molecules may be introduced during the sampling process (contamination) or may have been transported to the sampled environment by human or natural processes. Detection of these species should be reported and flagged or explained, when possible, to clarify the probable pathway into the sample. There are several ways DNA can be introduced, including transport by natural systems (Yamamoto et al. 2017; Mathieu et al. 2020; Darling et al. 2021; Baetscher et al. 2024). DNA moves downgradient through natural transport systems such as currents, surface runoff, or wind (Shogren et al. 2017; Harrison et al. 2019; Goray et al. 2024; Xiong et al. 2025). These detections are representative of the larger environmental system in the region, but may be misleading if presented in a list of organisms detected in a local ecosystem (Stolarz et al. 2024). It may be difficult or impossible to differentiate such detections

from DNA generated by more localized sources without robust sampling designs and known fate and transport within a given system (Xiong et al. 2025). Nonetheless, eDNA practitioners should be wary of such potential exogenous sources of eDNA. The use of robust sampling designs, pilot studies, and local knowledge is paramount in ensuring accurate interpretation of eDNA signals.

3) Process-based erroneous detections

Capture, handling, and interpretation of eDNA is influenced by a complex interplay of environmental, physiological, and analytical processes that can result in erroneous detections of species that are not present in the environment (Harrison et al. 2019). Contamination can occur at any stage of the sampling process, introducing DNA from a species that was not present at a site (Goldberg et al. 2016). The use of non-template controls (e.g., blanks) at each stage of the workflow (sampling through PCR) aids in the identification of contaminating species and associated samples. In some cases, contaminating species may also be present in the sampled environment and thus careful consideration should be given to how such samples should be reported (Hülpüsch et al. 2023; Fierer et al. 2025). Ambiguity in taxonomic assignments can also occur because of poor resolution of a given locus or because of incompleteness of reference databases, resulting in erroneous taxonomic assignments (Keck et al. 2023). Species with highly similar or even identical sequences in the amplicon region used for identification can be misidentified bioinformatically. As discussed above, different taxonomic classification algorithms can be more or less conservative in their approach to assigning taxonomy, potentially resulting in processed-based erroneous detections.

Addressing uncertainty - Filtering low counts, blanks

eDNA metabarcoding is a powerful method for detecting rare species, but low read counts in metabarcoding data present a challenge because it can be difficult to separate signal from noise (Hülpüsch et al. 2023). The relative penalty of a false-negative vs a false-positive detection will differ among projects, depending on the project objectives. Data filtering parameters can be selected to be more or less permissive based on the objectives of the study. For example, low read count detections can arise from low-level contamination or sequencing artifacts, even when best practices are applied (Deiner et al. 2017). False positive detections are likely to increase when rare taxa are targeted (Sepulveda et al. 2020), and researchers and managers may have to adjust tolerance for false positive detections if high sensitivity for rare taxa is desired (Darling et al. 2021). This emphasizes the importance of using complementary methods to confirm eDNA metabarcoding results, indicating expected detections of species of management concern (e.g., newly introduced species)(Gold et al. 2021).

Detecting contamination

Analyzing field, extraction, and PCR negative controls may help researchers assess the risk of erroneous detection through low-level contamination and artifacts (Goldberg et al. 2016). If negative controls have detectable concentrations of target DNA (e.g. gel electrophoresis ~ 10 ng limit of detection or ddPCR ~ 0.5

copies per reaction) then they must be sequenced to identify clear sources of contamination. Researchers may decide to sequence or ignore negative controls that do not obviously amplify target DNA (Van Der Loos and Nijland 2021). If blanks are sequenced, there should be a clear protocol indicating how these samples are interpreted and follow-up actions taken to address observed contamination.

Approaches for handling observed contamination include removing sequences observed in negative controls, identifying the pattern of exact ASVs (not just taxa) across multiple samples, sites, and controls; eliminating samples or ASVs based on the number of observed reads to remove noise from datasets (e.g., remove singletons or samples with fewer than 10,000 total reads), and comparing the similarity of biological and technical replicates. Many of these approaches can be implemented through existing software designed to detect and remove contamination, like the R programs Decontam (Davis et al. 2018) and metabar (Zinger et al. 2021). It is important to acknowledge the tradeoffs between more stringent thresholds used in decontamination processes (e.g. higher read count) which can reduce the risk of false positive detections, but which may also increase the risk of false negative detections. Researchers should be cognizant of the effects of such approaches on the interpretation of these results.

Comparing across replicate samples

Observing taxa in multiple replicates increases the confidence of detection (Royle 2006). Replicates of eDNA metabarcoding data can be different samples from the same time and place (same or different bottles), samples from the same place but different times, and multiple aliquots of the same sample. All increase confidence in results; if an organism is detected in more than 1 replicate, it may be less likely to be an artifact or introduced as contamination. Increasing the number of samples at any point in the workflow also increases costs, so the value of additional samples must be balanced against the resources required to process samples. In cases where a high degree of certainty is required, replicate samples may be essential.

Confidence versus resolution tradeoffs

As highlighted throughout, assigning species names to eDNA metabarcoding-derived sequences involves a broad suite of choices that may influence which species are reported. Previous work has documented that these choices result in confidence versus resolution tradeoffs, where choices result in either higher confidence decisions with more conservative species identification (e.g., genus or family rank) or less conservative species-level assignments with lower confidence. For example, previous work found that accuracy, overclassification, and misclassification rates increased with lower confidence score cutoffs in metabarcoding taxonomic assignment (Edgar 2018a; Gold et al. 2021). Thus, selection of the optimal parameter metric is contingent on what metric, accuracy maximization, bias minimization, or a combination of the two, is being optimized. Given the broad range of applications of eDNA approaches, differences in project objectives, and tolerances for inaccuracy, no one-size-fits-all approach is available for assigning taxonomy to eDNA sequences. Thus, it is contingent on the eDNA practitioners to discuss and identify relative tolerances and project objectives prior to project

implementation to inform confidence versus resolution tradeoffs. Many tradeoffs can be eliminated or at least ameliorated with additional background information, such as identifying priority target taxa, ensuring complete and comprehensive reference sequences, conducting *in silico* and *in vitro* validation and benchmarking of approaches, and delineating what trustworthiness of eDNA data results look like.

Implications of workflow process for species identification

Table 1 presents examples of commonly encountered issues related to species identification. Generally, each step in the processes described here has an impact on downstream steps and the resolution of species. Primer selection broadly

Table 1. Common issues encountered in species identification. Likely issues, and their causes are listed in Sector A, and the corresponding solution is shown in Sector B. For example, if non-native species are detected there are multiple possible reasons, ranging from the actual presence of that species in the environment, to errors in the data base. For each possible cause Sector B suggests possible solution.

Sector A. Issues that cause problems in species identification							
Issue	**Environmental**	**Contamination**	**Laboratory**		**Bioinformatics & Taxonomic Assignment**		
			Primer Choice	**PCR and Sequencing**	**Sequence QC & Dereplication**	**Classification**	**Reference Database**
Non-native species detected	Non-native species is present in sampled environment. Or DNA from anthropogenic sources present	Laboratory contamination in extraction or amplification step	Gene region cannot distinguish among native and non-native species	PCR artifact generates highly similar ASV to correct species / tag jumping from other samples on sequencing plate	QC not stringent enough to exclude artifacts	Classifier method accepts best match not perfect match.	Reference database lacks sequence from true species but contains identical sequence from non-native organism
Species-level taxonomy not identified			Primers do not amplify target species or gene region cannot distinguish among species		Dereplication collapses multiple true unique sequences (e.g. OTU clustering)	Selected classifier and/or parameters lack resolution to delineate species.	Multiple species have identical reference sequences
Species known to be present is not detected	Species is small component of sampled community and not captured in eDNA sampling		Primers do not amplify target species or gene region cannot distinguish among species	PCR amplification bias leads to lack of detection. Too few reads sequenced to capture full diversity.	QC too stringent and excludes true ASVs	Selected classifier and/or parameters lack resolution to delineate species.	Reference database lacks sequence from true species
Sector B. Solutions to problems listed in Sector A.							
Solutions	**Environmental**	**Contamination**	**Laboratory**		**Bioinformatics & Taxonomic Assignment**		
			Primer Choice	**PCR and Sequencing**	**Sequence QC & Dereplication**	**Classification**	**Reference Database**
Non-native species detected	Limit sampling to areas not impacted by known anthropogenic eDNA sources	Ensure proper sterile technique in field and laboratory work. Use multiple negative controls and monitor sequencing quality. Carefully track sample handling.	Use primers that have been validated for your target both in silico and in vitro.	Use methods like dual unique indices to limit tag jumping and DNA polymerases with low error rates.	Conduct stringent QC during pre and post processing. Apply abundance read thresholds or replicate-based filtering, informed by ecological plausibility.	Independently validate taxonomic assignments using other approaches. Provide "confidence flags" or a tiered approach to species reporting.	Ensure local reference database includes known non-native species. If needed generate additional reference sequences to enhance coverage.
Higher taxonomic ranks observed (e.g. Genus or Family level ID)			Use primers that have been validated for your target both in silico and in vitro.		Use unique ASVs without clustering as taxonomic identifiers	Conduct classifier and/or parameter optimization for maximum resolution. Perform sensitivity analyses on parameters.	Conduct taxonomic and geographic curation of reference databases.
Species known to be present is not detected	Sample higher volume | increase replication | use appropriate filter size		Use more specific primer to organism of interest and which amplifies region that can distinguish between closely related taxa	Increase sequencing depth/ replicates, adjust PCR parameters depending on direction of bias (towards or against target taxa)	Alter bioinformatic parameters to be more tolerant of rare sequences	Conduct classifier and/or parameter optimization to ensure maximum possible resolution. Limit reference database to local species.	Barcode local species and add to reference database. Include closely related global species in database and look for similar sequences.

constrains the organisms that are amplified, from domain (e.g., bacteria vs eukaryotes) to species-level resolution. PCR and sequencing steps must be carefully followed to prevent erroneous detections from contamination, or loss of species due to degradation. Bioinformatics encompasses multiple steps to filter, search, and assign taxonomy. Overly stringent criteria in these steps may remove taxa or reduce the ability to assign taxonomy to species level. Less stringent criteria may result in incorrect assignment of similar species. Finally, communication of results must clearly convey any uncertainties associated with taxonomic assignment, such as similar matches to multiple species, to avoid misunderstanding by end users.

Dissemination and Communication

All environmental monitoring methods have some underlying uncertainty; closely related species may be difficult to distinguish visually, taxonomy may be unresolved, net size or capture method may screen out some organisms, and very rare species may be essentially undetectable in some environments (Kelly et al. 2017). Monitoring applications relying on molecular methods have unique factors that influence species identification in ways that may be unfamiliar and require additional interpretation and end user support (Gold et al. 2021; Kelly et al. 2024). Careful consideration of the factors discussed above can reduce ambiguity, but species assignments from metabarcoding data will always involve some degree of uncertainty. Helping resource managers and other users understand the source and degree of this uncertainty is critical for responsible decision-making and to support trust in the data.

Thus, it's important to note and communicate the underlying uncertainty. Providing clear information on the methods and caveats associated with a given species identification helps users interpret the data correctly and decide on any necessary further steps. Researchers should also consider context-dependent reporting of eDNA metabarcoding observations. The choice between reporting a consensus taxonomic level or the overlapping species that comprise a taxonomic assignment depends on the specific requirements of the downstream analysis. If end users, such as resource managers or conservation biologists, require detailed species-level data, then it may be worth listing all possible candidate assignments for a given sequence. For fisheries management, conservation planning, or ecological studies, knowing the possible species, even with some uncertainty, can guide more targeted follow-up studies or interventions. In contexts where precise species-level information is critical (e.g., for detecting invasive species or ensuring sustainable fishing practices), providing the list of candidate species may be more informative.

Following FAIR and open science principles, scientists generating metabarcoding-based data sets must provide users with materials explaining how eDNA data is generated, and where sources of error come from if requested. Data reports should contain all information needed to understand the results, and we suggest that eDNA metabarcoding research papers for conservation and management should also include:

- A "Key Findings and Cautions" Section which summarizes major results, major results in plain language, but flags uncertainties clearly. For example, "Some detections (e.g., family *Gadus*) could not be reliably assigned

to species due to marker limitations. Follow-up visual surveys are recommended to confirm species identity where management action is needed."

- Quality control and quality assurance flags are used to indicate both the level and source of uncertainty. For example, if a sequence matches multiple closely related species, but only one is likely to be present, a researcher may choose to report only the taxonomy of the geographically relevant taxa. Here, the researcher should report the flag and when and where it was applied. Depending on project needs, users may choose to use the geographically relevant species-level assignment or rely only on genus-level assignment. Flags should indicate the level of confidence in the reported assignment and why that confidence is assigned, e.g., high confidence at species level, vs at genus only. Visualizations can help managers understand certainty levels.
- Explain when filtering criteria, such as minimum read counts or detection across replicates, were applied and how that affects results. These criteria may differ depending on project goals: when trying to detect a rare species, lower thresholds may be appropriate. All results should be available in some appendix or portion of the report to avoid filtering very rare species.
- Provide actionable guidance despite ambiguity. In many cases, data tables are generated by a contracted lab and cannot be tailored to meet specific project needs. In this case, general recommendations can be applied, such as reporting at genus level if resolution of an ambiguous species is not important to the study goals.

Conservation & management

As eDNA is increasingly used by a wider community of resource managers and other stakeholders, metabarcoding offers a powerful tool for detecting species presence and biodiversity in aquatic systems, but species identification from sequence data is influenced by a range of methodological and interpretive decisions. This framework outlines best practices for each stage of the workflow—primer selection, lab handling, bioinformatics, taxonomic assignment, and results communication—with a focus on transparency and aligning decisions with project objectives. By acknowledging limitations such as incomplete reference databases, contamination risks, ambiguous sequences, and detection biases, researchers can better interpret results and communicate findings to stakeholders. Replication, data validation, and the use of confidence flags or structured annotations can support responsible data use, especially in conservation, management, or regulatory contexts. Ultimately, this guidance can improve the accuracy, reliability, and usability of eDNA species identification while fostering trust and understanding among diverse end users.

Methods and approach used in this document

This paper focuses on portions of the metabarcoding workflow that are key to identifying species, but there are many additional, important aspects to metabarcoding-based ecosystem assessment that are not covered here. The use of eDNA to quantify organism abundance or population size, population genetics, and phenolic expression are emerging methodologies with high potential usefulness, but are not discussed here.

The document was conceived and written by all authors, with sections and figures provided by different authors. AI tools were used in the later stages of document preparation for final copy editing and reference management.

Acknowledgements

This is PMEL contribution number 5812. Zachary Gold was supported by NOAA's Pacific Marine Environmental Laboratory and the NOAA's Omics Program. We thank Dr. Sean McAllister and several anonymous reviewers for insightful feedback and suggestions. KJP was supported by the David and Lucile Packard Foundation. This work was supported by the U.S. Marine Biodiversity Observation Network (MBON) co-organized by NOAA, NASA, BOEM, and ONR through the National Oceanographic Partnership Program (NOPP).

Additional information

Conflict of interest

The authors have declared that no competing interests exist.

Ethical statement

No ethical statement was reported.

Use of AI

The following AI tools were used in the preparation of this manuscript: Chat GPT was used for final copy editing and reference managemen.

Funding

No funding was reported.

Author contributions

Conceptualization: AEH, KJP, EG, NP, ZG, KJ, AWW, KS. Methodology: NP, AEH, KS, EG, KJ, ZG, KJP. Project administration: AWW. Supervision: AWW. Writing - original draft: AT, KJP, FEMK, TK, ZG, LRT, AWW, KS, EM, JDB, SJ, EG, NP, AK, ML, NGA, FEB, KJ, AEH, JM. Writing - review and editing: ZG, ML, KJP, NGA, LRT, AT, EG, EM, MBO, AEH, NP, AWW, KJ, AK.

Author ORCIDs

AuthorName https://orcid.org/0000-0001-9700-6393
Alison W. Watts https://orcid.org/0000-0003-0490-7630
Zachary Gold https://orcid.org/0000-0001-8522-7682
Nastassia Patin https://orcid.org/0000-0003-1174-4243
Nicolaus G. Adams https://orcid.org/0000-0002-0673-7535
Jacoby D. Baker https://orcid.org/0000-0001-5204-6244
Fouad El Baidouri https://orcid.org/0000-0001-5204-6244
Erin Grey https://orcid.org/0000-0001-5883-0013
Ann E. Holmes https://orcid.org/0000-0003-4775-868X
Kym Jacobson https://orcid.org/0009-0003-5838-4174
Sean Jungbluth https://orcid.org/0000-0001-9265-8341
Tadashi Kajita https://orcid.org/0000-0003-1232-4179
Anders Kiledal https://orcid.org/0000-0003-4565-2187

Matthew Lemay https://orcid.org/0000-0001-7051-0020
Enrique Montes https://orcid.org/0000-0002-3103-4479
Frank E. Muller-Karger https://orcid.org/0000-0003-3159-5011
Jeffrey Miller https://orcid.org/0000-0003-2326-1855
Matthew B. Ogburn https://orcid.org/0000-0001-5417-555X
Kathleen J. Pitz https://orcid.org/0000-0002-4931-8592
Katherine Silliman https://orcid.org/0000-0001-5964-3965
Andrew Thompson https://orcid.org/0000-0002-1384-2229
Luke R. Thompson https://orcid.org/0000-0002-3911-1280

Data availability

All of the data that support the findings of this study are available in the main text or Supplementary Information.

References

Abarenkov K, Nilsson RH, Larsson K-H, Taylor AFS, May TW, Frøslev TG, Pawlowska J, Lindahl B, Põldmaa K, Truong C, Vu D, Hosoya T, Niskanen T, Piirmann T, Ivanov F, Zirk A, Peterson M, Cheeke TE, Ishigami Y, Jansson AT, Jeppesen TS, Kristiansson E, Mikryukov V, Miller JT, Oono R, Ossandon FJ, Paupério J, Saar I, Schigel D, Suija A, Tedersoo L, Kõljalg U (2024) The UNITE database for molecular identification and taxonomic communication of fungi and other eukaryotes: sequences, taxa and classifications reconsidered. Nucleic Acids Research 52: D791–D797. https://doi.org/10.1093/nar/gkad1039

Adams CIM, Hepburn C, Jeunen G, Cross H, Taylor HR, Gemmell NJ, Bunce M, Knapp M (2023) Environmental DNA reflects common haplotypic variation. Environmental DNA 5: 906–919. https://doi.org/10.1002/edn3.352

Ahmed M, Back MA, Prior T, Karssen G, Lawson R, Adams I, Sapp M (2019) Metabarcoding of soil nematodes: the importance of taxonomic coverage and availability of reference sequences in choosing suitable marker(s). Metabarcoding and Metagenomics 3: e36408. https://doi.org/10.3897/mbmg.3.36408

Alberdi A, Aizpurua O, Gilbert MTP, Bohmann K (2018) Scrutinizing key steps for reliable metabarcoding of environmental samples. Methods in Ecology and Evolution 9: 134–147. https://doi.org/10.1111/2041-210X.12849

Amaral-Zettler LA, McCliment EA, Ducklow HW, Huse SM (2009) A Method for Studying Protistan Diversity Using Massively Parallel Sequencing of V9 Hypervariable Regions of Small-Subunit Ribosomal RNA Genes. PLoS ONE 4: e6372. https://doi.org/10.1371/journal.pone.0006372

Amir A, McDonald D, Navas-Molina JA, Kopylova E, Morton JT, Zech Xu Z, Kightley EP, Thompson LR, Hyde ER, Gonzalez A, Knight R (2017) Deblur Rapidly Resolves Single-Nucleotide Community Sequence Patterns. mSystems 2: e00191-16. https://doi.org/10.1128/mSystems.00191-16

Anderson SR, Thompson LR (2022) Optimizing an enclosed bead beating extraction method for microbial and fish environmental DNA. Environmental DNA 4: 291–303. https://doi.org/10.1002/edn3.251

Anon (2020) The Global Biodiversity Information Facility. What is GBIF? https://www.gbif.org/what-is-gbif [January 5, 2026]

Anon (2026) OBIS: Ocean Biodiversity Information System, Intergovernmental Oceanographic Commission of UNESCO. OBIS. https://obis.org

Arranz V, Pearman WS, Aguirre JD, Liggins L (2020) MARES, a replicable pipeline and curated reference database for marine eukaryote metabarcoding. Scientific Data 7: 209. https://doi.org/10.1038/s41597-020-0549-9

Baetscher DS, Pochardt MR, Barry PD, Larson WA (2024) Tide impacts the dispersion of eDNA from nearshore net pens in a dynamic high-latitude marine environment. Environmental DNA 6: e533. https://doi.org/10.1002/edn3.533

Barbera P, Kozlov AM, Czech L, Morel B, Darriba D, Flouri T, Stamatakis A (2019) EPA-ng: Massively Parallel Evolutionary Placement of Genetic Sequences. Systematic Biology 68: 365–369. https://doi.org/10.1093/sysbio/syy054

Beng KC, Corlett RT (2020) Applications of environmental DNA (eDNA) in ecology and conservation: opportunities, challenges and prospects. Biodiversity and Conservation 29: 2089–2121. https://doi.org/10.1007/s10531-020-01980-0

Bengtsson-Palme J, Richardson RT, Meola M, Wurzbacher C, Tremblay ÉD, Thorell K, Kanger K, Eriksson KM, Bilodeau GJ, Johnson RM, Hartmann M, Nilsson RH (2018) Metaxa2 Database Builder: enabling taxonomic identification from metagenomic or metabarcoding data using any genetic marker. Bioinformatics 34: 4027–4033. https://doi.org/10.1093/bioinformatics/bty482

Benson DA, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW (2015) GenBank. Nucleic Acids Research 43: D30–D35. https://doi.org/10.1093/nar/gku1216

Bernatchez L, Ferchaud A-L, Berger CS, Venney CJ, Xuereb A (2024) Genomics for monitoring and understanding species responses to global climate change. Nature Reviews Genetics 25: 165–183. https://doi.org/10.1038/s41576-023-00657-y

Berry O, Jarman S, Bissett A, Hope M, Paeper C, Bessey C, Schwartz MK, Hale J, Bunce M (2021) Making environmental DNA (eDNA) biodiversity records globally accessible. Environmental DNA 3: 699–705. https://doi.org/10.1002/edn3.173

Blackman RC, Walser J, Rüber L, Brantschen J, Villalba S, Brodersen J, Seehausen O, Altermatt F (2023) General principles for assignments of communities from eDNA: Open versus closed taxonomic databases. Environmental DNA 5: 326–342. https://doi.org/10.1002/edn3.382

Bokulich NA, Kaehler BD, Rideout JR, Dillon M, Bolyen E, Knight R, Huttley GA, Gregory Caporaso J (2018) Optimizing taxonomic classification of marker-gene amplicon sequences with QIIME 2's q2-feature-classifier plugin. Microbiome 6: 90. https://doi.org/10.1186/s40168-018-0470-z

Bolyen E, Rideout JR, Dillon MR, Bokulich NA, Abnet CC, Al-Ghalith GA, Alexander H, Alm EJ, Arumugam M, Asnicar F, Bai Y, Bisanz JE, Bittinger K, Brejnrod A, Brislawn CJ, Brown CT, Callahan BJ, Caraballo-Rodríguez AM, Chase J, Cope EK, Da Silva R, Diener C, Dorrestein PC, Douglas GM, Durall DM, Duvallet C, Edwardson CF, Ernst M, Estaki M, Fouquier J, Gauglitz JM, Gibbons SM, Gibson DL, Gonzalez A, Gorlick K, Guo J, Hillmann B, Holmes S, Holste H, Huttenhower C, Huttley GA, Janssen S, Jarmusch AK, Jiang L, Kaehler BD, Kang KB, Keefe CR, Keim P, Kelley ST, Knights D, Koester I, Kosciolek T, Kreps J, Langille MGI, Lee J, Ley R, Liu Y-X, Loftfield E, Lozupone C, Maher M, Marotz C, Martin BD, McDonald D, McIver LJ, Melnik AV, Metcalf JL, Morgan SC, Morton JT, Naimey AT, Navas-Molina JA, Nothias LF, Orchanian SB, Pearson T, Peoples SL, Petras D, Preuss ML, Pruesse E, Rasmussen LB, Rivers A, Robeson MS, Rosenthal P, Segata N, Shaffer M, Shiffer A, Sinha R, Song SJ, Spear JR, Swafford AD, Thompson LR, Torres PJ, Trinh P, Tripathi A, Turnbaugh PJ, Ul-Hasan S, van der Hooft JJJ, Vargas F, Vázquez-Baeza Y, Vogtmann E, von Hippel M, Walters W, Wan Y, Wang M, Warren J, Weber KC, Williamson CHD, Willis AD, Xu ZZ, Zaneveld JR, Zhang Y, Zhu

Q, Knight R, Caporaso JG (2019) Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nature Biotechnology 37: 852–857. https://doi.org/10.1038/s41587-019-0209-9

Boratyn GM, Camacho C, Cooper PS, Coulouris G, Fong A, Ma N, Madden TL, Matten WT, McGinnis SD, Merezhuk Y, Raytselis Y, Sayers EW, Tao T, Ye J, Zaretskaya I (2013) BLAST: a more efficient report with usability improvements. Nucleic Acids Research 41: W29–W33. https://doi.org/10.1093/nar/gkt282

Bourlat SJ, Haenel Q, Finnman J, Leray M (2016) Preparation of Amplicon Libraries for Metabarcoding of Marine Eukaryotes Using Illumina MiSeq: The Dual-PCR Method. In: Bourlat SJ (Ed.) Marine Genomics. Methods in Molecular Biology. Springer New York, New York, NY, 197–207. https://doi.org/10.1007/978-1-4939-3774-5_13

Boyer F, Mercier C, Bonin A, Le Bras Y, Taberlet P, Coissac E (2016) OBITOOLS: a UNIX -inspired software package for DNA metabarcoding. Molecular Ecology Resources 16: 176–182. https://doi.org/10.1111/1755-0998.12428

Brandão-Dias PF, Shaffer M, Guri G, Parsons KM, Kelly RP, Allan EA (2025) Differential decay of multiple environmental nucleic acid components. Scientific Reports 15: 26791. https://doi.org/10.1038/s41598-025-12916-5

Bucklin A, Lindeque PK, Rodriguez-Ezpeleta N, Albaina A, Lehtiniemi M (2016) Metabarcoding of marine zooplankton: prospects, progress and pitfalls. Journal of Plankton Research 38: 393–400. https://doi.org/10.1093/plankt/fbw023

Buetas E, Jordán-López M, López-Roldán A, D'Auria G, Martínez-Priego L, De Marco G, Carda-Diéguez M, Mira A (2024) Full-length 16S rRNA gene sequencing by PacBio improves taxonomic resolution in human microbiome samples. BMC Genomics 25: 310. https://doi.org/10.1186/s12864-024-10213-5

Callahan BJ, McMurdie PJ, Holmes SP (2017) Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. The ISME Journal 11: 2639–2643. https://doi.org/10.1038/ismej.2017.119

Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP (2016) DADA2: High-resolution sample inference from Illumina amplicon data. Nature Methods 13: 581–583. https://doi.org/10.1038/nmeth.3869

Callahan BJ, Wong J, Heiner C, Oh S, Theriot CM, Gulati AS, McGill SK, Dougherty MK (2019) High-throughput amplicon sequencing of the full-length 16S rRNA gene with single-nucleotide resolution. Nucleic Acids Research 47: e103–e103. https://doi.org/10.1093/nar/gkz569

Çevik T, Çevik N (2025) Environmental DNA (eDNA): A review of ecosystem biodiversity detection and applications. Biodiversity and Conservation 34: 2999–3035. https://doi.org/10.1007/s10531-025-03112-y

Chiarello M, McCauley M, Villéger S, Jackson CR (2022) Ranking the biases: The choice of OTUs vs. ASVs in 16S rRNA amplicon data analysis has stronger effects on diversity measures than rarefaction and OTU identity threshold. PLoS ONE 17: e0264443. https://doi.org/10.1371/journal.pone.0264443

Chorlton SD (2024) Ten common issues with reference sequence databases and how to mitigate them. Frontiers in Bioinformatics 4: 1278228. https://doi.org/10.3389/fbinf.2024.1278228

Clarke LJ, Beard JM, Swadling KM, Deagle BE (2017) Effect of marker choice and thermal cycling protocol on zooplankton DNA metabarcoding studies. Ecology and Evolution 7: 873–883. https://doi.org/10.1002/ece3.2667

Collins RA, Bakker J, Wangensteen OS, Soto AZ, Corrigan L, Sims DW, Genner MJ, Mariani S (2019) Non-specific amplification compromises environmental DNA metabarcoding with COI. Methods in Ecology and Evolution 10: 1985–2001. https://doi.org/10.1111/2041-210X.13276

Curd EE, Gal L, Gallego R, Silliman K, Nielsen S, Gold Z (2024) rCRUX: A rapid and versatile tool for generating metabarcoding reference libraries in R. Environmental DNA 6: e489. https://doi.org/10.1002/edn3.489

Curd EE, Gold Z, Kandlikar GS, Gomer J, Ogden M, O'Connell T, Pipes L, Schweizer TM, Rabichow L, Lin M, Shi B, Barber PH, Kraft N, Wayne R, Meyer RS (2019) Anacapa Toolkit: An environmental DNA toolkit for processing multilocus metabarcode datasets. Methods in Ecology and Evolution 10: 1469–1475. https://doi.org/10.1111/2041-210X.13214

Curto M, Veríssimo A, Riccioni G, Santos CD, Ribeiro F, Jentoft S, Alves MJ, Gante HF (2025) Improving Whole Biodiversity Monitoring and Discovery With Environmental DNA Metagenomics. Molecular Ecology Resources 25: e14105. https://doi.org/10.1111/1755-0998.14105

Darling JA, Jerde CL, Sepulveda AJ (2021) What do you mean by false positive? Environmental DNA 3: 879–883. https://doi.org/10.1002/edn3.194

Davis NM, Proctor DM, Holmes SP, Relman DA, Callahan BJ (2018) Simple statistical identification and removal of contaminant sequences in marker-gene and metagenomics data. Microbiome 6: 226. https://doi.org/10.1186/s40168-018-0605-2

Deiner K, Bik HM, Mächler E, Seymour M, Lacoursière-Roussel A, Altermatt F, Creer S, Bista I, Lodge DM, de Vere N, Pfrender ME, Bernatchez L (2017) Environmental DNA metabarcoding: Transforming how we survey animal and plant communities. Molecular Ecology 26: 5872–5895. https://doi.org/10.1111/mec.14350

DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, Huber T, Dalevi D, Hu P, Andersen GL (2006) Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Applied and Environmental Microbiology 72: 5069–5072. https://doi.org/10.1128/AEM.03006-05

Dickie IA, Boyer S, Buckley HL, Duncan RP, Gardner PP, Hogg ID, Holdaway RJ, Lear G, Makiola A, Morales SE, Powell JR, Weaver L (2018) Towards robust and repeatable sampling methods in eDNA -based studies. Molecular Ecology Resources 18: 940–952. https://doi.org/10.1111/1755-0998.12907

Dos Santos RA, Blabolil P (2025) Comparison of Bioinformatic Pipelines for eDNA Metabarcoding Data Analysis of Fish Populations. Fishes 10: 214. https://doi.org/10.3390/fishes10050214

Duarte S, Simões L, Costa FO (2023) Current status and topical issues on the use of eDNA-based targeted detection of rare animal species. Science of The Total Environment 904: 166675. https://doi.org/10.1016/j.scitotenv.2023.166675

Dunthorn M, Lipps JH, Dolan JR, Saab MA-A, Aescht E, Bachy C, De Cao MSB, Berger H, Bourland WA, Choi JK, Clamp J, Doherty M, Gao F, Gentekaki E, Gong J, Hu X, Huang J, Kamiyama T, Johnson MD, Kammerlander B, Kim SY, Kim Y-O, La Terza A, Laval-Peuto M, Lipscomb D, Lobban CS, Long H, Luporini P, Lynn DH, Macek M, Mansergh RI, Martín-Cereceda M, McManus GG, Montagnes DJS, Ong'ondo GO, Patterson DJ, Pérez-Uz B, Quintela-Alonso P, Safi LSL, Santoferrara LF, Sonntag B, Song W, Stoeck T, Stoecker DK, Strüder-Kypke MC, Trautmann I, Utz LRP, Vallesi A, Vd'ačný P, Warren A, Weisse T, Wickham SA, Yi Z, Zhang W, Zhan Z, Zufall R, Agatha S (2015) Ciliates — Protists with complex morphologies and ambiguous early fossil record. Marine Micropaleontology 119: 1–6. https://doi.org/10.1016/j.marmicro.2015.05.004

Edgar RC (2018a) Accuracy of taxonomy prediction for 16S rRNA and fungal ITS sequences. PeerJ 6: e4652. https://doi.org/10.7717/peerj.4652

Edgar RC (2018b) Updating the 97% identity threshold for 16S ribosomal RNA OTUs. Bioinformatics 34: 2371–2375. https://doi.org/10.1093/bioinformatics/bty113

Edgar RC, Flyvbjerg H (2015) Error filtering, pair assembly and error correction for next-generation sequencing reads. Bioinformatics 31: 3476–3482. https://doi.org/10.1093/bioinformatics/btv401

El Baidouri F, Watts AW, Miller JT, Kelly M, Sevigny JL, Gilbert H, Thomas WK (2025) An optimized eDNA protocol for fish tracking in estuarine environments. Scientific Reports 15: 1175. https://doi.org/10.1038/s41598-025-85176-y

Espinosa Prieto A, Hardion L, Debortoli N, Beisel J (2024) Finding the perfect pairs: A matchmaking of plant markers and primers for multi-marker eDNA metabarcoding. Molecular Ecology Resources 24: e13937. https://doi.org/10.1111/1755-0998.13937

Ferreira AO, Azevedo OM, Barroso C, Duarte S, Egas C, Fontes JT, Ré P, Santos AMP, Costa FO (2024) Multi-marker DNA metabarcoding for precise species identification in ichthyoplankton samples. Scientific Reports 14: 19772. https://doi.org/10.1038/s41598-024-69963-7

Fierer N, Leung PM, Lappan R, Eisenhofer R, Ricci F, Holland SI, Dragone N, Blackall LL, Dong X, Dorador C, Ferrari BC, Goordial J, Holmes SP, Inagaki F, Korem T, Li SS, Makhalanyane TP, Metcalf JL, Nagarajan N, Orsi WD, Shanahan ER, Walker AW, Weyrich LS, Gilbert JA, Willis AD, Callahan BJ, Shade A, Parkhill J, Banfield JF, Greening C (2025) Guidelines for preventing and reporting contamination in low-biomass microbiome studies. Nature Microbiology 10: 1570–1580. https://doi.org/10.1038/s41564-025-02035-2

Fontes JT, Katoh K, Pires R, Soares P, Costa FO (2024) Benchmarking the discrimination power of commonly used markers and amplicons in marine fish (e)DNA (meta) barcoding. Metabarcoding and Metagenomics 8: e128646. https://doi.org/10.3897/mbmg.8.128646

Freeland JR (2017) The importance of molecular markers and primer design when characterizing biodiversity from environmental DNA. Genome 60: 358–374. https://doi.org/10.1139/gen-2016-0100

Frøslev TG, Kjøller R, Bruun HH, Ejrnæs R, Brunbjerg AK, Pietroni C, Hansen AJ (2017) (cite) Algorithm for post-clustering curation of DNA amplicon data yields reliable biodiversity estimates. Nature Communications 8: 1188. https://doi.org/10.1038/s41467-017-01312-x

Galaska M, Brown S, McAllister S (2023) Monitoring Biodiversity Impacts of a Changing Arctic Through Environmental DNA. Oceanography. https://doi.org/10.5670/oceanog.2023.221

Gamage G, Gimhana N, Wickramarachchi A, Mallawaarachchi V, Perera I (2019) Alignment-free Whole Genome Comparison Using k-mer Forests. In: 2019 19th International Conference on Advances in ICT for Emerging Regions (ICTer), 1–7. https://doi.org/10.1109/ICTer48817.2019.9023714

Gohl DM, Vangay P, Garbe J, MacLean A, Hauge A, Becker A, Gould TJ, Clayton JB, Johnson TJ, Hunter R, Knights D, Beckman KB (2016) Systematic improvement of amplicon marker gene methods for increased accuracy in microbiome studies. Nature Biotechnology 34: 942–949. https://doi.org/10.1038/nbt.3601

Gold Z, Curd EE, Goodwin KD, Choi ES, Frable BW, Thompson AR, Walker Jr HJ, Burton RS, Kacev D, Martz LD, Barber PH (2021) Improving metabarcoding taxonomic as-

signment: A case study of fishes in a large marine ecosystem. Molecular Ecology Resources 21: 2546–2564. https://doi.org/10.1111/1755-0998.13450

Gold Z, Wall AR, Schweizer TM, Pentcheff ND, Curd EE, Barber PH, Meyer RS, Wayne R, Stolzenbach K, Prickett K, Luedy J, Wetzer R (2022) A manager's guide to using eDNA metabarcoding in marine ecosystems. PeerJ 10: e14071. https://doi.org/10.7717/peerj.14071

Goldberg CS, Turner CR, Deiner K, Klymus KE, Thomsen PF, Murphy MA, Spear SF, McKee A, Oyler-McCance SJ, Cornman RS, Laramie MB, Mahon AR, Lance RF, Pilliod DS, Strickler KM, Waits LP, Fremier AK, Takahara T, Herder JE, Taberlet P (2016) Critical considerations for the application of environmental DNA methods to detect aquatic species. Methods in Ecology and Evolution 7: 1299–1307. https://doi.org/10.1111/2041-210X.12595

Goray M, Taylor D, Bibbo E, Fantinato C, Fonneløp AE, Gill P, van Oorschot RAH (2024) Emerging use of air eDNA and its application to forensic investigations - A review. Electrophoresis 45: 916–932. https://doi.org/10.1002/elps.202300228

Guillou L, Bachar D, Audic S, Bass D, Berney C, Bittner L, Boutte C, Burgaud G, de Vargas C, Decelle J, del Campo J, Dolan JR, Dunthorn M, Edvardsen B, Holzmann M, Kooistra WHCF, Lara E, Le Bescot N, Logares R, Mahé F, Massana R, Montresor M, Morard R, Not F, Pawlowski J, Probert I, Sauvadet A-L, Siano R, Stoeck T, Vaulot D, Zimmermann P, Christen R (2013) The Protist Ribosomal Reference database (PR2): a catalog of unicellular eukaryote Small Sub-Unit rRNA sequences with curated taxonomy. Nucleic Acids Research 41: D597–D604. https://doi.org/10.1093/nar/gks1160

Hakimzadeh A, Abdala Asbun A, Albanese D, Bernard M, Buchner D, Callahan B, Caporaso JG, Curd E, Djemiel C, Brandström Durling M, Elbrecht V, Gold Z, Gweon HS, Hajibabaei M, Hildebrand F, Mikryukov V, Normandeau E, Özkurt E, M Palmer J, Pascal G, Porter TM, Straub D, Vasar M, Větrovský T, Zafeiropoulos H, Anslan S (2024) A pile of pipelines: An overview of the bioinformatics software for metabarcoding data analyses. Molecular Ecology Resources 24: e13847. https://doi.org/10.1111/1755-0998.13847

Harrison JB, Sunday JM, Rogers SM (2019) Predicting the fate of eDNA in the environment and implications for studying biodiversity. Proceedings of the Royal Society B: Biological Sciences 286: 20191409. https://doi.org/10.1098/rspb.2019.1409

Hastings PA, Burton RS (2008) Establishing a DNA Sequence Database for the Marine Fish Fauna of California. https://escholarship.org/uc/item/1ck9b3qs [January 7, 2026]

Hleap JS, Littlefair JE, Steinke D, Hebert PDN, Cristescu ME (2021) Assessment of current taxonomic assignment strategies for metabarcoding eukaryotes. Molecular Ecology Resources 21: 2190–2203. https://doi.org/10.1111/1755-0998.13407

Hülpüsch C, Rauer L, Nussbaumer T, Schwierzeck V, Bhattacharyya M, Erhart V, Traidl-Hoffmann C, Reiger M, Neumann AU (2023) Benchmarking MicrobIEM – a user-friendly tool for decontamination of microbiome sequencing data. BMC Biology 21: 269. https://doi.org/10.1186/s12915-023-01737-5

Huse SM, Welch DM, Morrison HG, Sogin ML (2010) Ironing out the wrinkles in the rare biosphere through improved OTU clustering. Environmental Microbiology 12: 1889–1898. https://doi.org/10.1111/j.1462-2920.2010.02193.x

Illumina (n.d.) PhiX Control v3. https://www.illumina.com/products/by-type/sequencing-kits/cluster-gen-sequencing-reagents/phix-control-v3.html [accessed 15 November 2025]

Irber L, Pierce-Ward NT, Abuelanin M, Alexander H, Anant A, Barve K, Baumler C, Botvinnik O, Brooks P, Dsouza D, Gautier L, Hera MR, Houts HE, Johnson LK, Klötzl F, Koslicki D, Lim M, Lim R, Nelson B, Ogasawara I, Reiter T, Scott C, Sjödin A, Standage

D, Swamidass SJ, Tiffany C, Vemuri P, Young E, Brown CT (2024) sourmash v4: A multitool to quickly search, compare, and analyze genomic and metagenomic data sets. Journal of Open Source Software 9: 6830. https://doi.org/10.21105/joss.06830

Ishii R, Poti T, Akimitsu K, Narumi-Kawasaki T, Yamada Y (2025) Species-specific detection of environmental DNA by endpoint PCR using mismatch base introducing primer design and nested PCR. Limnology 26: 253–261. https://doi.org/10.1007/s10201-025-00786-9

Iwasaki W, Fukunaga T, Isagozawa R, Yamada K, Maeda Y, Satoh TP, Sado T, Mabuchi K, Takeshima H, Miya M, Nishida M (2013) MitoFish and MitoAnnotator: A Mitochondrial Genome Database of Fish with an Accurate and Automatic Annotation Pipeline. Molecular Biology and Evolution 30: 2531–2540. https://doi.org/10.1093/molbev/mst141

Jeunen G, Lamare M, Cummings V, Treece J, Ferreira S, Massuger J, Pryor Rodgers L, Tait L, Lust B, Wilkinson S, Mariani S, Mills S, Gemmell N (2023a) Unveiling the hidden diversity of marine eukaryotes in the Ross Sea: A comparative analysis of seawater and sponge eDNA surveys. Environmental DNA 5: 1780–1792. https://doi.org/10.1002/edn3.500

Jeunen G-J, Dowle E, Edgecombe J, von Ammon U, Gemmell NJ, Cross H (2023b) crabs—A software program to generate curated reference databases for metabarcoding sequencing data. Molecular Ecology Resources 23: 725–738. https://doi.org/10.1111/1755-0998.13741

Kaehler BD, Bokulich NA, McDonald D, Knight R, Caporaso JG, Huttley GA (2019) Species abundance information improves sequence taxonomy classification accuracy. Nature Communications 10: 4643. https://doi.org/10.1038/s41467-019-12669-6

Kalendar R, Shevtsov A, Otarbay Z, Ismailova A (2024) In silico PCR analysis: a comprehensive bioinformatics tool for enhancing nucleic acid amplification assays. Frontiers in Bioinformatics 4: 1464197. https://doi.org/10.3389/fbinf.2024.1464197

Keck F, Couton M, Altermatt F (2023) Navigating the seven challenges of taxonomic reference databases in metabarcoding analyses. Molecular Ecology Resources 23: 742–755. https://doi.org/10.1111/1755-0998.13746

Kelly RP, Shelton AO, Gallego R (2019) Understanding PCR Processes to Draw Meaningful Conclusions from Environmental DNA Studies. Scientific Reports 9: 12133. https://doi.org/10.1038/s41598-019-48546-x

Kelly RP, Closek CJ, O'Donnell JL, Kralj JE, Shelton AO, Samhouri JF (2017) Genetic and Manual Survey Methods Yield Different and Complementary Views of an Ecosystem. Frontiers in Marine Science 3. https://doi.org/10.3389/fmars.2016.00283

Kelly RP, Lodge DM, Lee KN, Theroux S, Sepulveda AJ, Scholin CA, Craine JM, Andruszkiewicz Allan E, Nichols KM, Parsons KM, Goodwin KD, Gold Z, Chavez FP, Noble RT, Abbott CL, Baerwald MR, Naaum AM, Thielen PM, Simons AL, Jerde CL, Duda JJ, Hunter ME, Hagan JA, Meyer RS, Steele JA, Stoeckle MY, Bik HM, Meyer CP, Stein E, James KE, Thomas AC, Demir-Hilton E, Timmers MA, Griffith JF, Weise MJ, Weisberg SB (2024) Toward a national eDNA strategy for the United States. Environmental DNA 6: e432. https://doi.org/10.1002/edn3.432

Kiledal EA, Reitz LA, Kuiper EQ, Evans J, Siddiqui R, Denef VJ, Dick GJ (2024) Comparative genomic analysis of Microcystis strain diversity using conserved marker genes. Harmful Algae 132: 102580. https://doi.org/10.1016/j.hal.2024.102580

Klymus KE, Baker JD, Abbott CL, Brown RJ, Craine JM, Gold Z, Hunter ME, Johnson MD, Jones DN, Jungbluth MJ, Jungbluth SP, Lor Y, Maloy AP, Merkes CM, Noble R, Patin NV, Sepulveda AJ, Spear SF, Steele JA, Takahashi M, Watts AW, Theroux S (2024) The MIEM guidelines: Minimum information for reporting of environmental metabarcoding data. Metabarcoding and Metagenomics 8: e128689. https://doi.org/10.3897/mbmg.8.128689

Kozich JJ, Westcott SL, Baxter NT, Highlander SK, Schloss PD (2013) Development of a Dual-Index Sequencing Strategy and Curation Pipeline for Analyzing Amplicon Sequence Data on the MiSeq Illumina Sequencing Platform. Applied and Environmental Microbiology 79: 5112–5120. https://doi.org/10.1128/AEM.01043-13

Ledger KJ, Hicks MBR, Hurst TP, Larson W, Baetscher DS (2024) Validation of Environmental DNA for Estimating Proportional and Absolute Biomass. Environmental DNA 6: e70030. https://doi.org/10.1002/edn3.70030

Ledger KJ, Everett M, Nichols KM, Larson WA, Baetscher DS (2025) Development and Validation of an eDNA Primer Set for Rockfishes in the Alaska Current Provides a New Tool for Fisheries Management. Environmental DNA 7: e70176. https://doi.org/10.1002/edn3.70176

Leray M, Yang JY, Meyer CP, Mills SC, Agudelo N, Ranwez V, Boehm JT, Machida RJ (2013) A new versatile primer set targeting a short fragment of the mitochondrial COI region for metabarcoding metazoan diversity: application for characterizing coral reef fish gut contents. Frontiers in Zoology 10: 34. https://doi.org/10.1186/1742-9994-10-34

Lu J, Salzberg SL (2020) Ultrafast and accurate 16S rRNA microbial community analysis using Kraken 2. Microbiome 8: 124. https://doi.org/10.1186/s40168-020-00900-2

Lyet A, Pellissier L, Valentini A, Dejean T, Hehmeyer A, Naidoo R (2021) eDNA sampled from stream networks correlates with camera trap detection rates of terrestrial mammals. Scientific Reports 11: 11362. https://doi.org/10.1038/s41598-021-90598-5

Macher T-H, Schütz R, Yildiz A, Beermann AJ, Leese F (2023) Evaluating five primer pairs for environmental DNA metabarcoding of Central European fish species based on mock communities. Metabarcoding and Metagenomics 7: e103856. https://doi.org/10.3897/mbmg.7.103856

Mahon AR, Grey EK, Jerde CL (2022) Integrating invasive species risk assessment into environmental metabarcoding reference libraries. Ecological Applications 33: e2730. https://doi.org/10.1002/eap.2730

Marín A, Alfaro R, Reyes-Flores LE, Ingar C, Santos-Rojas LE, Alvarez-Jaque IB, Rodríguez-Bernales K, Carbajal C, Yon-Utrilla A, Zelada-Mázmela E (2025) Development and validation of versatile species-specific primer assays for eDNA monitoring and authentication of 10 commercially important Peruvian marine species. PLOS One 20: e0313181. https://doi.org/10.1371/journal.pone.0313181

Marques V, Milhau T, Albouy C, Dejean T, Manel S, Mouillot D, Juhel J-B (2021) (cite) GAPeDNA: Assessing and mapping global species gaps in genetic databases for eDNA metabarcoding. Diversity and Distributions 27: 1880–1892. https://doi.org/10.1111/ddi.13142

Mathieu C, Hermans SM, Lear G, Buckley TR, Lee KC, Buckley HL (2020) A Systematic Review of Sources of Variability and Uncertainty in eDNA Data for Environmental Monitoring. Frontiers in Ecology and Evolution Volume 8-2020. https://www.frontiersin.org/journals/ecology-and-evolution/articles/10.3389/fevo.2020.00135

Mathon L, Valentini A, Guérin P, Normandeau E, Noel C, Lionnet C, Boulanger E, Thuiller W, Bernatchez L, Mouillot D, Dejean T, Manel S (2021) Benchmarking bioinformatic tools for fast and accurate eDNA metabarcoding species identification. Molecular Ecology Resources 21: 2565–2579. https://doi.org/10.1111/1755-0998.13430

Matsen FA, Kodner RB, Armbrust EV (2010) pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. BMC Bioinformatics 11: 538. https://doi.org/10.1186/1471-2105-11-538

Min MA, Barber PH, Gold Z (2021) MiSebastes: An eDNA metabarcoding primer set for rockfishes (genus Sebastes). Conservation Genetics Resources 13: 447–456. https://doi.org/10.1007/s12686-021-01219-2

Miya M, Sato Y, Fukunaga T, Sado T, Poulsen JY, Sato K, Minamoto T, Yamamoto S, Yamanaka H, Araki H, Kondoh M, Iwasaki W (2015) MiFish, a set of universal PCR primers for metabarcoding environmental DNA from fishes: detection of more than 230 subtropical marine species. Royal Society Open Science 2: 150088. https://doi.org/10.1098/rsos.150088

Mugnai F, Costantini F, Chenuil A, Leduc M, Gutiérrez Ortega JM, Meglécz E (2023) Be positive: customized reference databases and new, local barcodes balance false taxonomic assignments in metabarcoding studies. PeerJ 11: e14616. https://doi.org/10.7717/peerj.14616

Nilsson RH, Bok G, Ryberg M, Kristiansson E, Hallenberg N (2009) A software pipeline for processing and identification of fungal ITS sequences. Source Code for Biology and Medicine 4: 1. https://doi.org/10.1186/1751-0473-4-1

O'Brien TD, Blanco-Bercial L, Questel JM, Batta-Lona PG, Bucklin A (2024) MetaZoo-Gene Atlas and Database: Reference Sequences for Marine Ecosystems. Methods in Molecular Biology (Clifton, N.J.) 2744: 475–489. https://doi.org/10.1007/978-1-0716-3581-0_28

Ogbanga N, Nelson A, Gino S, Wescott DJ, Mickleburgh HL, Gocha TP, Procopio N (2023) The impact of freezing on the post-mortem human microbiome. Frontiers in Ecology and Evolution 11: 1151001. https://doi.org/10.3389/fevo.2023.1151001

Pagenkopp Lohan KM, Aguilar R, DiMaria R, Heggie K, Tuckey TD, Fabrizio MC, Ogburn MB (2023) Juvenile Striped Bass consume diverse prey in Chesapeake Bay tributaries. Marine and Coastal Fisheries 15: e210259. https://doi.org/10.1002/mcf2.10259

Parada AE, Needham DM, Fuhrman JA (2016) Every base matters: assessing small subunit rRNA primers for marine microbiomes with mock communities, time series and global field samples. Environmental Microbiology 18: 1403–1414. https://doi.org/10.1111/1462-2920.13023

Parks DH, Chaumeil P-A, Mussig AJ, Rinke C, Chuvochina M, Hugenholtz P (2025) GTDB release 10: a complete and systematic taxonomy for 715 230 bacterial and 17 245 archaeal genomes. Nucleic Acids Research: gkaf1040. https://doi.org/10.1093/nar/gkaf1040

Parks DH, Chuvochina M, Waite DW, Rinke C, Skarshewski A, Chaumeil P-A, Hugenholtz P (2018) A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. Nature Biotechnology 36: 996–1004. https://doi.org/10.1038/nbt.4229

Patin NV, Goodwin KD (2023) Capturing marine microbiomes and environmental DNA: A field sampling guide. Frontiers in Microbiology 13: 1026596. https://doi.org/10.3389/fmicb.2022.1026596

Peterson DL, Allen MC, Vastano A, Lockwood JL (2022) Evaluation of sample collection and storage protocols for surface eDNA surveys of an invasive terrestrial insect. Environmental DNA 4: 1201–1211. https://doi.org/10.1002/edn3.314

Petit-Marty N, Casas L, Saborido-Rey F (2023) State-of-the-art of data analyses in environmental DNA approaches towards its applicability to sustainable fisheries management. Frontiers in Marine Science 10: 1061530. https://doi.org/10.3389/fmars.2023.1061530

Pipes L, Nielsen R (2024) A rapid phylogeny-based method for accurate community profiling of large-scale metabarcoding datasets. eLife 13: e85794. https://doi.org/10.7554/eLife.85794

Polanco FA, Richards E, Flück B, Valentini A, Altermatt F, Brosse S, Walser J, Eme D, Marques V, Manel S, Albouy C, Dejean T, Pellissier L (2021) Comparing the performance of 12S mitochondrial primers for fish environmental DNA across ecosystems. Environmental DNA 3: 1113–1127. https://doi.org/10.1002/edn3.232

Poulsen CS, Kaas RS, Aarestrup FM, Pamp SJ (2021) Standard Sample Storage Conditions Have an Impact on Inferred Microbiome Composition and Antimicrobial Resistance Patterns. Microbiology Spectrum 9: e01387-21. https://doi.org/10.1128/Spectrum.01387-21

Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glöckner FO (2013) The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Research 41: D590–D596. https://doi.org/10.1093/nar/gks1219

Quince C, Lanzen A, Davenport RJ, Turnbaugh PJ (2011) Removing Noise From Pyrosequenced Amplicons. BMC Bioinformatics 12: 38. https://doi.org/10.1186/1471-2105-12-38

Ratnasingham S, Wei C, Chan D, Agda J, Agda J, Ballesteros-Mejia L, Boutou HA, El Bastami ZM, Ma E, Manjunath R, Rea D, Ho C, Telfer A, McKeowan J, Rahulan M, Steinke C, Dorsheimer J, Milton M, Hebert PDN (2024) BOLD v4: A Centralized Bioinformatics Platform for DNA-Based Biodiversity Data. Methods in Molecular Biology (Clifton, N.J.) 2744: 403–441. https://doi.org/10.1007/978-1-0716-3581-0_26

Renner HM, Piatt JF, Renner M, Drummond BA, Laufenberg JS, Parrish JK (2024) Catastrophic and persistent loss of common murres after a marine heatwave. Science 386: 1272–1276. https://doi.org/10.1126/science.adq4330

Richardson RT, Bengtsson-Palme J, Johnson RM (2017) Evaluating and optimizing the performance of software commonly used for the taxonomic classification of DNA metabarcoding sequence data. Molecular Ecology Resources 17: 760–769. https://doi.org/10.1111/1755-0998.12628

Rimet F, Gusev E, Kahlert M, Kelly MG, Kulikovskiy M, Maltsev Y, Mann DG, Pfannkuchen M, Trobajo R, Vasselon V, Zimmermann J, Bouchez A (2019) Diat.barcode, an open-access curated barcode library for diatoms. Scientific Reports 9: 15116. https://doi.org/10.1038/s41598-019-51500-6

Robeson MS, O'Rourke DR, Kaehler BD, Ziemski M, Dillon MR, Foster JT, Bokulich NA (2021) RESCRIPt: Reproducible sequence taxonomy reference database management. PLOS Computational Biology 17: e1009581. https://doi.org/10.1371/journal.pcbi.1009581

Rognes T, Flouri T, Nichols B, Quince C, Mahé F (2016) VSEARCH: a versatile open source tool for metagenomics. PeerJ 4: e2584. https://doi.org/10.7717/peerj.2584

Royle JA (2006) Site occupancy models with heterogeneous detection probabilities. Biometrics 62: 97–102. https://doi.org/10.1111/j.1541-0420.2005.00439.x

Sales NG, Wangensteen OS, Carvalho DC, Deiner K, Præbel K, Coscia I, McDevitt AD, Mariani S (2021) Space-time dynamics in monitoring neotropical fish communities using eDNA metabarcoding. Science of The Total Environment 754: 142096. https://doi.org/10.1016/j.scitotenv.2020.142096

Schoch CL, Ciufo S, Domrachev M, Hotton CL, Kannan S, Khovanskaya R, Leipe D, Mcveigh R, O'Neill K, Robbertse B, Sharma S, Soussov V, Sullivan JP, Sun L, Turner S, Karsch-Mizrachi I (2020) NCBI Taxonomy: a comprehensive update on curation, resources and tools. Database 2020: baaa062. https://doi.org/10.1093/database/baaa062

Sepulveda AJ, Hutchins PR, Forstchen M, Mckeefry MN, Swigris AM (2020) The Elephant in the Lab (and Field): Contamination in Aquatic Environmental DNA Studies. Frontiers in Ecology and Evolution 8: 609973. https://doi.org/10.3389/fevo.2020.609973

Serrana JM, Watanabe K (2023) Haplotype-level metabarcoding of freshwater macroinvertebrate species: A prospective tool for population genetic analysis. PLOS ONE 18: e0289056. https://doi.org/10.1371/journal.pone.0289056

Shaffer MR, Andruszkiewicz Allan E, Van Cise AM, Parsons KM, Shelton AO, Kelly RP (2025) Observation Bias in Metabarcoding. Molecular Ecology Resources 25: e14119. https://doi.org/10.1111/1755-0998.14119

Shea MM, Kuppermann J, Rogers MP, Smith DS, Edwards P, Boehm AB (2023) Systematic review of marine environmental DNA metabarcoding studies: toward best practices for data usability and accessibility. PeerJ 11: e14993. https://doi.org/10.7717/peerj.14993

Shogren AJ, Tank JL, Andruszkiewicz E, Olds B, Mahon AR, Jerde CL, Bolster D (2017) Controls on eDNA movement in streams: Transport, Retention, and Resuspension. Scientific Reports 7: 5065. https://doi.org/10.1038/s41598-017-05223-1

Singer GAC, Fahner NA, Barnes JG, McCarthy A, Hajibabaei M (2019) Comprehensive biodiversity analysis via ultra-deep patterned flow cell technology: a case study of eDNA metabarcoding seawater. Scientific Reports 9: 5991. https://doi.org/10.1038/s41598-019-42455-9

Singh MG, Wahengbam R (2025) Optimization of denoising and filtering parameters of DADA2 for QIIME2 amplicon metagenomics data analysis, 1–20. https://doi.org/10.1101/2025.06.24.661437

So KYK, Fong JJ, Lam IPY, Dudgeon D (2020) Pitfalls during in silico prediction of primer specificity for eDNA surveillance. Ecosphere 11: e03193. https://doi.org/10.1002/ecs2.3193

Spens J, Evans AR, Halfmaerten D, Knudsen SW, Sengupta ME, Mak SST, Sigsgaard EE, Hellström M (2017) Comparison of capture and storage methods for aqueous macrobial eDNA using an optimized extraction protocol: advantage of enclosed filter. Yu D (Ed.) Methods in Ecology and Evolution 8: 635–645. https://doi.org/10.1111/2041-210X.12683

Srinivas M, Walsh CJ, Crispie F, O'Sullivan O, Cotter PD, Van Sinderen D, Kenny JG (2025) Evaluating the efficiency of 16S-ITS-23S operon sequencing for species level resolution in microbial communities. Scientific Reports 15: 2822. https://doi.org/10.1038/s41598-024-83410-7

Stein ED, Jerde CL, Allan EA, Sepulveda AJ, Abbott CL, Baerwald MR, Darling J, Goodwin KD, Meyer RS, Timmers MA, Thielen PM (2024) Critical considerations for communicating environmental DNA science. Environmental DNA 6: e472. https://doi.org/10.1002/edn3.472

Stoeckle MY, Ausubel JH, Coogan M (2024a) 12S gene metabarcoding with DNA standard quantifies marine bony fish environmental DNA, identifies threshold for reproducible detection, and overcomes distortion due to amplification of non-fish DNA. Environmental DNA 6: e376. https://doi.org/10.1002/edn3.376

Stoeckle MY, Ausubel JH, Hinks G, VanMorter SM (2024b) A potential tool for marine biogeography: eDNA-dominant fish species differ among coastal habitats and by season concordant with gear-based assessments. PLoS ONE 19: e0313170. https://doi.org/10.1371/journal.pone.0313170

Stoeckle MY, Adolf J, Ausubel JH, Charlop-Powers Z, Dunton KJ, Hinks G (2022) Current laboratory protocols for detecting fish species with environmental DNA optimize sensitivity and reproducibility, especially for more abundant populations. ICES Journal of Marine Science 79: 403–412. https://doi.org/10.1093/icesjms/fsab273

Stolarz T, Henley B, Roarty H, Kohut J, Adolf J (2024) Determining the Origin and Fate of Marine eDNA. In: 2024 IEEE/OES Thirteenth Current, Waves and Turbulence

Measurement (CWTM). IEEE, Wanchese, NC, USA, 1–6. https://doi.org/10.1109/CWTM61020.2024.10526336

Takahashi M, Saccò M, Kestel JH, Nester G, Campbell MA, Van Der Heyde M, Heydenrych MJ, Juszkiewicz DJ, Nevill P, Dawkins KL, Bessey C, Fernandes K, Miller H, Power M, Mousavi-Derazmahalleh M, Newton JP, White NE, Richards ZT, Allentoft ME (2023) Aquatic environmental DNA: A review of the macro-organismal biomonitoring revolution. Science of The Total Environment 873: 162322. https://doi.org/10.1016/j.scitotenv.2023.162322

Takahashi M, Frøslev TG, Paupério J, Thalinger B, Klymus K, Helbing CC, Villacorta-Rath C, Silliman K, Thompson LR, Jungbluth SP, Yong SY, Formel S, Jenkins G, Laporte M, Deagle B, Rajbhandari S, Jeppesen TS, Bissett A, Jerde C, Hahn EE, Schriml LM, Hunter C, Newman P, Woollard P, Harper LR, Dunn N, West K, Haderlé R, Wilkinson S, Acharya-Patel N, Lopez MLD, Cochrane G, Berry O (2025) A Metadata Checklist and Data Formatting Guidelines to Make eDNA FAIR (Findable, Accessible, Interoperable, and Reusable). Environmental DNA 7: e70100. https://doi.org/10.1002/edn3.70100

Theroux S, Sepulveda A, Abbott CL, Gold Z, Watts AW, Hunter ME, Klymus KE, Hirsch SL, Craine JM, Jones DN, Brown RJ, Steele JA, Takahashi M, Noble RT, Darling JA (2025) What is eDNA method standardisation and why do we need it? Metabarcoding and Metagenomics 9: e132076. https://doi.org/10.3897/mbmg.9.132076

Tressoldi PE (2011) Extraordinary Claims Require Extraordinary Evidence: The Case of Non-Local Perception, a Classical and Bayesian Review of Evidences. Frontiers in Psychology 2. https://doi.org/10.3389/fpsyg.2011.00117

Van Der Loos LM, Nijland R (2021) Biases in bulk: DNA metabarcoding of marine communities and the methodology involved. Molecular Ecology 30: 3270–3288. https://doi.org/10.1111/mec.15592

Van Etten J, Stephens TG, Bhattacharya D (2023) A *k*- mer-Based Approach for Phylogenetic Classification of Taxa in Environmental Genomic Data. Systematic Biology 72: 1101–1118. https://doi.org/10.1093/sysbio/syad037

Vandepitte L, Vanhoorne B, Decock W, Dekeyzer S, Trias Verbeeck A, Bovit L, Hernandez F, Mees J (2015) How Aphia—The Platform behind Several Online and Taxonomically Oriented Databases—Can Serve Both the Taxonomic Community and the Field of Biodiversity Informatics. Journal of Marine Science and Engineering 3: 1448–1473. https://doi.org/10.3390/jmse3041448

Vasselon V, Rivera SF, Ács É, Almeida SFP, Andree KB, Apothéloz-Perret-Gentil L, Bailet B, Baričević A, Beentjes KK, Bettig J, Bouchez A, Capelli C, Chardon C, Duleba M, Elersek T, Genthon C, Jablonska M, Jacas L, Kahlert M, Kelly MG, Macher J-N, Mauri F, Moletta-Denat M, Mortágua A, Pawlowski J, Pérez-Burillo J, Pfannkuchen M, Pilgrim E, Pissaridou P, Rimet F, Stanic K, Tapolczai K, Theroux S, Trobajo R, Van Der Hoorn B, Vasquez MI, Vidal M, Wanless D, Warren J, Zimmermann J, Paix B (2025) Proficiency testing and cross-laboratory method comparison to support standardisation of diatom DNA metabarcoding for freshwater biomonitoring. Metabarcoding and Metagenomics 9: e133264. https://doi.org/10.3897/mbmg.9.133264

Vinje H, Liland KH, Almøy T, Snipen L (2015) Comparing K-mer based methods for improved classification of 16S sequences. BMC Bioinformatics 16: 205. https://doi.org/10.1186/s12859-015-0647-4

Wang Q, Garrity GM, Tiedje JM, Cole JR (2007) Naïve Bayesian Classifier for Rapid Assignment of rRNA Sequences into the New Bacterial Taxonomy. Applied and Environmental Microbiology 73: 5261–5267. https://doi.org/10.1128/AEM.00062-07

Wood DE, Salzberg SL (2014) Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biology 15: R46. https://doi.org/10.1186/gb-2014-15-3-r46

Xiong J, MacCready P, Brasseale E, Andruszkiewicz Allan E, Ramón-Laca A, Parsons KM, Shaffer M, Kelly RP (2025) Advective Transport Drives Environmental DNA Dispersal in an Estuary. Environmental Science & Technology 59: 7506–7516. https://doi.org/10.1021/acs.est.5c01286

Xu Y, Huo X, Chen X, Wang Z, Zhou M, Zhu J, Yan R, Cai Y (2024) Comparing the performance variability of different eDNA primers in fish monitoring. Watershed Ecology and the Environment 6: 165–175. https://doi.org/10.1016/j.wsee.2024.08.003

Yamamoto S, Masuda R, Sato Y, Sado T, Araki H, Kondoh M, Minamoto T, Miya M (2017) Environmental DNA metabarcoding reveals local fish communities in a species-rich coastal sea. Scientific Reports 7: 40368. https://doi.org/10.1038/srep40368

Zhang S, Zhao J, Yao M (2020) A comprehensive and comparative evaluation of primers for metabarcoding eDNA from fish. Methods in Ecology and Evolution 11: 1609–1625. https://doi.org/10.1111/2041-210X.13485

Zhu T, Iwasaki W (2023) MultiBarcodeTools: Easy selection of optimal primers for eDNA multi-metabarcoding. Environmental DNA 5: 1793–1808. https://doi.org/10.1002/edn3.499

Zhu X, Bell KL, Rourke ML, Wu H, Gopurenko D (2024) Generic qPCR assays for quality control in environmental DNA research. Environmental DNA 6: e536. https://doi.org/10.1002/edn3.536

Zinger L, Lionnet C, Benoiston A, Donald J, Céline Mercier, Boyer F (2021) metabaR: An R package for the evaluation and improvement of DNA metabarcoding data quality. Methods in Ecology and Evolution 12: 586–592. https://doi.org/10.1111/2041-210X.13552

Zizka A, Antunes Carvalho F, Calvente A, Rocio Baez-Lizarazo M, Cabral A, Coelho JFR, Colli-Silva M, Fantinati MR, Fernandes MF, Ferreira-Araújo T, Gondim Lambert Moreira F, Santos NMC, Santos TAB, Dos Santos-Costa RC, Serrano FC, Alves Da Silva AP, De Souza Soares A, Cavalcante De Souza PG, Calisto Tomaz E, Vale VF, Vieira TL, Antonelli A (2020) No one-size-fits-all solution to clean GBIF. PeerJ 8: e9916. https://doi.org/10.7717/peerj.9916

Supplementary material 1

Examples of current reference databases

Authors: Alison W. Watts, Zachary Gold, Nastassia Patin, Nicolaus G. Adams, Jacoby D. Baker, Fouad El Baidouri, Erin Grey, Ann E. Holmes, Kym Jacobson, Sean Jungbluth, Tadashi Kajita, Anders Kiledal, Matthew Lemay, Enrique Montes, Frank E. Muller-Karger, Jeffrey Miller, Matthew B. Ogburn, Kathleen J. Pitz, Katherine Silliman, Andrew Thompson, Luke R. Thompson

Data type: xlsx

Explanation note: Examples of reference databases used in metabarcoding species identification. This is not comprehensive, but indicates the type of data available.

Copyright notice: This dataset is made available under the Open Database License (http://opendatacommons.org/licenses/odbl/1.0/). The Open Database License (ODbL) is a license agreement intended to allow users to freely share, modify, and use this Dataset while maintaining this same freedom for others, provided that the original source and author(s) are credited.

Link: https://doi.org/10.3897/mbmg.10.174450.suppl1